Teaching medical ethics

Evaluating ethics competence in medical education

Julian Savulescu, Roger Crisp, K W M Fulford and Tony Hope The Murdoch Institute, Melbourne, Australia, St Anne’s College, Oxford University, University of Warwick and University of Oxford

Abstract

We critically evaluate the ways in which competence in medical ethics has been evaluated. We report the initial stage in the development of a relevant, reliable and valid instrument to evaluate core critical thinking skills in medical ethics. This instrument can be used to evaluate the impact of medical ethics education programmes and to assess whether medical students have achieved a satisfactory level of performance of core skills and knowledge in medical ethics, within and across institutions.

(Journal of Medical Ethics 1999;25:367–374)

Keywords: Medical ethics; medical education; evaluation; reliability; validity

I Background

THE DEVELOPMENT OF MEDICAL ETHICS TEACHING

There have been calls for the inclusion of more humanistic ideals in medical teaching for 20 years. One distinguishing characteristic of poor clinical performance is not only poor medical knowledge, but moral failing, including negligence in addressing the concerns of patients. Sheehan and colleagues' describe a “consistent and meaningful relationship” between level of moral reasoning and clinical performance. We believe that patient care could be improved not only by improvements in knowledge of the technical and scientific aspects of medicine, but also by improving the ethical sensitivities and analytic abilities of doctors.

Besides promoting more effective medical practice, medical ethics is important in itself. Many medical decisions have a moral component. Training in medical ethics involves developing the skills of logical moral reasoning. Medical ethics has thus become a standard component of many medical curricula. Most medical schools in North America have a medical humanities programme. Teaching of undergraduates in the clinical years is increasing, as is that of clinicians in practice. Physicians who have had formal teaching in medical ethics tend to perceive those courses to be of value. Medical students express a need for more, and better, ethics education.

In the UK, the General Medical Council has stated that medical ethics should constitute one of the components of the core medical curriculum. Recently, a Consensus Group of Teachers of Medical Ethics and Law in UK Medical Schools has described a detailed 12-point core curriculum for medical ethics teaching. This document sets out for the first time in the UK the core content of the curriculum for review; it outlines what is expected of new doctors, and it facilitates systematic and reliable evaluation across medical schools. It does not, however, describe how this core curriculum is to be taught, and in particular the level of skills students must display. Adoption and implementation of the core ethics curriculum is voluntary. In this paper, we argue that adoption of a curriculum has significant advantages.

In the University of Oxford clinical school, medical ethics is taught as a part of the Practice Skills Course. This was originally set up by Tony Hope and William Fulford, in association with Theo Schofield and William Rosenberg, to provide training in ethics, medical law and communications skills. A manual has been produced which provides detailed course materials, information on resources, and notes on conducting seminars and assessment. It is based on five years of experience in setting up a course in medical ethics.

THE PLACE OF EVALUATION OF COMPETENCE IN MEDICAL ETHICS

Increasing value is being placed on the humanistic and caring qualities of physicians: several North American bodies test for competence in medical ethics in their qualifying examinations.
medical faculties regard failure in medical ethics in the same way as failure in science or clinical subjects and require satisfactory performance in order to sit final examinations. In the UK, the Pond Report recommended that there should be compulsory assessment of medical students’ ability to reason through ethical issues. The development of methods of assessing performance in ethics has been described as the “next stage” following the development of a consensus statement on core curriculum in the UK.

Four reasons

Standardised evaluation of competence in ethics is important for at least four reasons. First, it is necessary for formal empirical study of medical ethics teaching and the modes of delivering it. Medical ethics teaching is a behavioural intervention intended to bring about certain desirable outcomes. There have been calls for formal evaluation of whether ethics teaching has any impact on the behaviour of clinicians.

The second reason is more broadly educational. One major barrier to the introduction of the Practice Skills Project in Oxford was the belief that “you can’t teach ethics until you have shown that your programme is effective”. One way to justify access to curriculum time is to demonstrate that ethics is a teachable subject that can be formally evaluated. Just as clinicians should be held accountable for the clinical competence of graduating doctors, so too medical ethicists should be accountable for their ethical competence.

Those responsible for medical courses quite rightly “seek assurance that the investment of scarce educational resources will result in a satisfactory and measurable dividend”. Formal evaluation brings to the surface what is being taught, for scrutiny and review and most importantly makes transparent the aims of medical ethics. It is also reinforces to students that medical ethics is as important as other medical disciplines.

Thirdly, while adoption of the core curriculum by medical faculties is voluntary, widespread employment of a core curriculum would make it easier to evaluate students across institutions. However, this also requires a reliable evaluative instrument.

The fourth reason why evaluation is important is that there is a two-way process between refining the core curriculum of knowledge and skills, and their evaluation. The question: “what is medical ethics” can be answered in part by how we evaluate it.

A significant proportion of medical students believe that ethics teaching should be “rigorous, required and examined”. Indeed, examination is thought to focus students’ minds.

CURRENT METHODS OF EVALUATING MEDICAL ETHICS TEACHING, THEIR RESULTS, AND THEIR RELIABILITY/VALIDITY

Attempts to develop standardised tests began with the moral judgment interview of Lawrence Kohlberg, based on his descriptive theory of moral development. This test involves a 30-minute semi-structured interview in which subjects are presented with three moral dilemmas and asked a set of open-ended questions. Kohlberg’s work has spawned other evaluative instruments. For example, Self, Wolinsky and Baldwin employed the socio-moral reflection measure of Gibbs, modified Kohlberg test, and showed that medical ethics teaching can increase the capacity for moral reasoning. Small-group case studies may be more effective than lectures (see also).

Other instruments to evaluate ethical sensitivity have been developed. Some involve subjectivity and reports. The most popular have employed clinical vignettes, after Kohlberg, followed by open-probe questions. Videotaped vignettes have also been used. Rest’s defining issues test employs vignettes followed by a number of responses. Respondents are asked to rank the importance of the response on a five-point scale.

Any instrument designed to evaluate competence in medical ethics must be reliable (ie, it must give the same score to the same student performance regardless of who employs it or when it is used) and valid (ie, it must actually measure ethical competence and not something else). Rest’s is probably the most extensively studied test of moral competence. There have been extensive tests of its reliability, including test-retest reliability (ie the ability of the test to give the same score when repeated at a later time) and internal coherence, but since it employs a standard menu of responses, issues of inter-rater reliability do not arise. (A test has inter-rater reliability if different raters employing it give the same score to the same student performance using the test.)

Vignette-based instrument

Hébert and colleagues used a vignette-based instrument to evaluate medical ethics teaching to different undergraduate years, evaluating their sensitivity (as ability to identify) to issues of autonomy, beneficence and justice. They claimed that ethical sensitivity decreases in the later part of undergraduate medical school.
colleagues observed a similar decay in competence. Hébert and colleagues claim to correlate ethical sensitivity with a career choice for family medicine; Sulmasy and colleagues found that ethical competence was greater in older Jewish doctors.

The scoring system for marking responses of Hébert and colleagues was based on a consensus of five experts. Reliability of the initial instrument was evaluated (Pearson coefficients by pairs of observers ranged from 0.89 to 0.94). The authors admit that the experts missed issues that were identified subsequently by students. There was no attempt to evaluate how reliable other raters would be using their marking system nor were there formal tests of validity.

In a randomised controlled trial of two different educational strategies for residents, Sulmasy and colleagues measured three variables: knowledge, confidence and responsiveness. This instrument consists of a 15-point multiple choice test of knowledge (which included questions concerning such issues as whether Maryland had a living will statute, the content of the Hippocratic Oath and the definition of utilitarianism) and an evaluation of confidence. There was also a case of futility which involved a patient with terminal cancer, with a ruptured oesophagus admitted in respiratory failure. Choices were: (1) intubation; (2) elicit the patient's preferences for options; (3) admit, provide oxygen, antipyretics and morphine.

Modest effects
The knowledge test was evaluated by two ethicists for face validity and inter-item reliability, and the confidence scale was evaluated for internal consistency. Sulmasy and colleagues showed that ethics education for residents may increase confidence in dealing with ethical problems in those residents who had perceived themselves as having had a poor undergraduate ethics education, but it had no impact on knowledge of ethics in most residents. In response to this, the educational programme was altered to include regular didactic conferences over a two-year period. Using a slightly modified evaluative instrument, this resulted in an increase in confidence from 3.3 to 3.8 on a five-point Likert scale (p<0.001) and a 14% increase in knowledge (p<0.001). The authors admit that the effects were "modest". The instrument tested content knowledge, not skills or reasoning.

At the University of Newcastle, Australia, Mitchell and colleagues test students' ethical competence using a modified essay-question, based on a case vignette. Model answers are written by those teaching the course. The assessment procedure must be approved by a committee comprising the year co-ordinator and the chairpersons of each academic term. Following assessment, and prior to marking, students review the assessment. Assessment aims to evaluate sensitivity to ethical issues, knowledge and ability to reflect critically. Problems identified with this mode of assessment include: student interpretation of the question; the number and weighting of the ethical elements in the model answers; and the level which represents the mandatory level of competence (the minimum necessary for a pass). There have been no formal tests of reliability or validity.

Doyal employs a case-vignette approach to evaluate medical ethics teaching at St Bartholomew's and the London Hospital Medical Colleges. Evaluation is according to a model answer. No test of reliability within the panel of markers is reported.

More ambitious modes of examining ethical performance are being proposed. Mitchell and colleagues have proposed an action-oriented assessment tool using an objective structured clinical examination (OSCE).

Evaluation of medical ethics teaching has suffered from a number of methodological problems. While it is not possible to evaluate each instrument in this paper, we believe that the following constitute some of the properties of an instrument for evaluating competence in medical ethics and some of the limitations of current approaches:

1. Any instrument for evaluating ethical knowledge and competence must be relevant. That is, it must measure outcomes that capture core knowledge and skills of the discipline - content validity. There are no statistical measures of validity in the above instruments. (Validity is most commonly assessed by asking experts to examine an instrument and assess whether it evaluates the right sorts of things. This could be done prospectively: experts could be asked to list the relevant features of an instrument "blind" to the instrument, and the statistical correlation between what experts believe is important and what is evaluated in the instrument assessed. That would be an example of a statistical estimate of validity.)

2. The instrument must be capable of being reliably applied by different raters. There have been some measures of inter-rater reliability, but no measures of naive rater reliability, that is, whether raters not involved in the development of the instrument are able to use the instrument reliably.
3. The instrument must be relevant to clinical practice.
4. The structure of the instrument must be publicly justifiable and subject to review.

II Aim: Developing a relevant, reliable and valid measure of critical thinking skills
We identified two core skills central to medical ethics: (i) being able to identify ethical issues (awareness) and (ii) being able to reason about these issues and come to a justifiable resolution as to what should be done. We called this second group of skills, critical thinking skills. We report the first stage in the development of a relevant, reliable and valid instrument for evaluating the second of critical thinking skills.

III Methods
Six vignettes were developed to evaluate thinking skills in ethics (an example of one vignette is found in appendix 1). Vignettes were constructed to reflect ethical issues arising in clinical practice.

A sample of these vignettes have been administered to two cohorts of students:

1. One hundred and sixty students; 80 in the first clinical year and 80 in the third clinical year (final year). Scripts from these students were used to test the instrument.
2. Eighty students from the second clinical year (prefinal year), repeated after two months to evaluate test-retest reliability. These students had no relevant ethics teaching during this period. The period was chosen to be long enough to ensure a low chance of remembering the previous response, but short enough to ensure that there was no relevant further learning. Scripts from these students were used to test the instrument.

In cohort 1, each student answered questions relating to one vignette. In cohort 3, students answered two vignettes. Each student received one vignette in common with the previous test. Students were given strict instructions to devote 7.5 minutes to the awareness vignette and 15 minutes to the thinking skill vignettes.

DEVELOPING THE MARKING SYSTEM
A marking system for evaluating the thinking skills was developed in the following way. Three markers were chosen who had formal training in philosophy or bioethics and were involved in medical ethics teaching (two were practising doctors): RC, KWMF and TH. Each marked six scripts for each vignette and met to look for areas of agreement and disagreement. All of these scripts came from cohort 1. A set of principles/criteria for marking were developed. A further six scripts were marked, and the same process repeated, refining the evaluative instrument. This was iterated two further times until a set of principles/criteria were agreed upon for marking all vignettes. This constitutes the marking instrument (an example of marking instructions appears in appendix 1). The face validity of the instrument was evaluated by these three experts.

After development of the instrument, the following were evaluated:

1. The inter-rater reliability of the instrument.
2. The content validity of the instrument.
3. Test-retest reliability.

Sixty scripts were randomly selected from cohort 2 (test-retest group). These consisted of 30 scripts from the initial group (time 1), and 30 from the same students retested two months later (time 2). There were five scripts for each vignette at each time. Within each vignette group, scripts were randomly selected by a person not involved in the study and presented to examiners blinded to student name.

Each of the three primary markers (RC, KWMF and TH) marked these scripts using the marking instrument. These marks were evaluated for inter-rater reliability. (That is, did the three primary raters give similar scores to each script?) Content validity was evaluated in the following way. A naïve marker (JS) was chosen, who, until this point had played no part in development of the instrument. Julian Savulescu had formal training in medicine and bioethics, and was involved in medical ethics education. He marked the scripts without using the instrument (blinded to the instrument). He was instructed simply to give each script a score between 0 and 10 reflecting its quality. The concordance of these marks with the marks of the three primary markers using the instrument was used as an evaluation of the content validity of the instrument as an assessment of core ethical thinking skills.

Test-retest reliability involved evaluating the extent to which the same student answering the same script two months later was given the same mark, from the same rater. Test-retest reliability was evaluated by comparing the 30 scores awarded by each of the three primary markers for each script in cohort 2 at time 1 with the 30 scores awarded to the scripts from the same student who responded to the same vignette two months later (time 2).
Results

Figure 1 shows the frequency with which each rater used each score. The median scores are within 0.5 mark of each other. The 25-75% interquartile range for all markers is within the range of marks from 2-4.

“Absolute differences in scores” by pairwise comparison of examiners (figure 2) shows a similar pattern for all. There were very few papers in which there were more than two marks difference. The majority of papers were within 0-1 mark of each other.

Spearman’s rank correlation showed a significant correlation between pairwise comparisons of all three markers (KWMF/RC p = 0.0017; KWMF/TH p = 0; RC/TH p = 0). Wilcoxon signed-rank tests showed that there were no significant differences between the medians (KWMF/RC p = 0.11; RC/TH p = 0.92), except between KWMF and TH (p = 0.01). Paired t-tests revealed no significant differences between the means (KWMF/RC p = 0.07; RC/TH p = 0.74), except between KWMF and TH (p = 0.01; 95% CI 0.08-0.65).

Content validity of the instrument: Naive marker (JS)

Figure 3 shows score distribution for the “naive marker” in comparison with the other raters. The distribution of marks is similar, but the median of
the naïve rater is four marks higher. The spread is similar, though extreme lower values are lower, probably as a result of the naïve rater operating at the upper rather than the lower end of the scale.

Figure 4 is a smoothed histogram showing the frequency of scores around the mean for that marker. The distribution of the naïve marker is very similar to that of the other three markers. Julian Savulescu is virtually identical to TH and RC, with KWMF having less spread. Julian Savulescu has a few relatively lower values, but this is a function of the fact that he is operating at the upper end of the range. Figure 5 shows the absolute differences between the scores of the naïve rater and each of the three primary raters. The negative values are an artefact of the statistical analysis used and should be ignored. Figure 5 shows a similar behaviour of JS compared to each of the three primary raters.

Spearman rank correlation pairwise between JS and the other three markers shows a significant correlation between JS/KWMF (p = 0.0034) and JS/TH (p = 0.012), but not between JS/RC (p = 0.0645), though the result is at the border of significance. The statistically significant correlation between the naïve rater and at least two of the three primary raters indicates that the marking instrument (as used by these markers) is valid.

**TEST-RETEST RELIABILITY**

If we assume that scores are normally distributed, analysis of variance shows that there is no significant difference between scores individually: students were given at time 1 compared with their student’s score at time 2 (Pr(F) = 0.84). This suggests good test-retest reliability. There was also no significant difference among the scores awarded by each of three different markers to individual scripts (Pr(F) = 0.15), which confirms inter-rater reliability.

**V Discussion**

On the basis of these results, this instrument provides a good assessment of ethical thinking skills. Raters showed good inter-rater reliability and there was good test-retest reliability. On the basis of the high correlation in scores between the blinded naïve expert rater and those using the instrument, it is a valid assessment of core skills. This is, to our knowledge, the first prospective blinded evaluation of the validity of such an instrument which ascribes a statistical level of confidence to the evaluation.

We believe that the instrument is a useful and relevant formal assessment of ethical thinking skills. However, an alternative interpretation of the data is that the instrument is redundant: it performs similarly to the naïve rater’s global assessment—indeed, this is our evidence for validity. The advantages of the instrument are, first, it articulates precisely why a student received a particular score; second, it has been shown to be reliable; third, it can potentially be used by less expert markers. To test whether the instrument is redundant, we would need to expose other naïve markers to the instrument, and then training in bioethics, to examine whether the instrument changed their marking behaviour. If the instrument is of value, we would expect it to change the behaviour of at least those markers who do not have training in bioethics.

More importantly, it remains to be seen whether raters, particularly without training in medical ethics, can apply the instrument reliably:

**Naive rater reliability.** The question of naïve rater reliability goes to the heart of scientific investigation. Many tests of ethical competence employ case vignettes which invite subjects to reason through a line of argument, demonstrating ethical awareness and knowledge. Investigators have attempted to evaluate rater reliability through the construction of a model answer. However, it remains open how raters will apply these criteria to individual answers. Indeed, problems of application have been identified. Some investigators have recognised this problem, and...
evaluated inter-rater reliability. However, this has been evaluated using those raters involved in the development of the instrument. This is a biased evaluation. If different naïve raters will award different scores when given the same answers to ethical questions and applying the same evaluative instrument, then utility of the instrument is questionable. In the educational setting, the problem of inter-rater reliability raises questions of fairness. These problems are not unique to medical ethics assessment. They are common to the evaluation of all clinical skills. We hope formally to evaluate naïve rater reliability in the future.

This instrument is the first stage in development of a comprehensive programme of evaluation of critical analytic skills in medical ethics. The development and application of such instruments will be enhanced by the adoption of a core curriculum in medical ethics teaching.

Appendix 1. Critical thinking skills vignettes

Here is one example of the vignettes with marking instructions.

During your morning general practice surgery a 50-year-old woman comes to see you, complaining of some mild clumsiness and worsening memory. She tells you that her father and his mother both died from Huntington's disease in their mid-fifties. The rest of her family do not know of the occurrence of Huntington's disease in her relatives. She insists that you do not tell them and forbids you to let them know her present concern. Later that morning this woman's daughter sees you to discuss coming off the contraceptive pill as she and her husband wish to conceive a child.

[There follows some information about Huntington's disease.]

What should you say to the daughter?
Justify your decision and note what other options are open to you.

Why do you think that these other options are less satisfactory than the one you chose to adopt?

Give one mark only for each point—broadly construed—made from the list below. Note any points made which you consider appropriate but which do not appear on the list.

1. Obligations and confidentiality.
2. "Implied contract" argument for respecting confidentiality.
3. Utilitarian argument for respecting confidentiality.

5. Duty of care to one's own patients.
7. Autonomy versus beneficence conflict.
8. Consequences for the potential child.
9. Interests of the husband.
10. Counselling of the mother.
11. Counselling of the daughter.
12. Advising the daughter to research into grandparents' medical history.

Acknowledgements

We would like to acknowledge the help of the following people: Drs Ian Eastwood and Jonathan Price for carrying out ratings, Dr Cortina for statistical advice, Professor Donald McIntyre for help with research design and Susanna Morey, Sean Weaver, Eleanor Korendow, Maryna Lewinski and Susannah Brock for help in collecting data. We also thank all the students who took part in the research.

Julian Savulescu, MB, BS, BmedSci, PhD, is Director of the Ethics Unit, The Murdoch Institute, Royal Children's Hospital, Melbourne, Australia and Associate Professor, the Bioethics Programme, Centre for the Study of Health and Society, University of Melbourne. Roger Crisp is Fellow and Tutor, St Anne's College, Oxford. K W M Fullford is Professor of Philosophy and Mental Health, University of Warwick and Honorary Consultant Psychiatrist, University of Oxford. Tony Hope is Director, The Oxford Institute for Ethics and Communication in Health Care Practice, University of Oxford Division of Public Health and Primary Care and Reader in Medicine, University of Oxford.

References

24 Medical Ethics Subcommittee of the American Board of Paediatrics. Teaching and evaluation of interpersonal skills and ethical decision making in paediatrics. Paediatrics 1987;79:829-34.
43 Komiyama AA. A workshop at the College of Medicine, Lagos University, Journal of Medical Ethics 1984;10:199-200.

News and notes

Journal of Medical Ethics - http://jmedethics.com

Visitors to the world wide web can now access the Journal of Medical Ethics either through the BMJ Publishing Group’s home page (http://www.bmjpg.com) or directly by using its individual URL (http://www.jmedethics.com). There they will find the following:

- Current contents list for the journal
- Contents lists of previous issues
- Members of the editorial board
- Journal information
- Instructions for authors
- Details of reprint services.