an extra economic burden on carers? Is social function maintained better because of hospital at home? Currently we are evaluating hospital at home care in these terms.

We thank the managements of Peterborough District Hospital and the hospital at home service for their support in this study, especially the respective finance departments for supplying the costing information. We are also grateful to members of departments within the district hospital for describing the service they provide, in particular the staff of the orthopaedic wards.

1 Boyce WJ, Vessey MP. Rising incidence of fracture of the proximal femur. *Lancet* 1985;i:150-1.
2 Central Statistical Office. *Social trends.* Vol 22. London: HMSO, 1992:27.
3 Campion EW, Jette AM, Cleary PD, Harris BA. Hip fracture: a prospective study of hospital course, complications and cost. *J Gen Intern Med* 1987;2:78-82.
4 Borgquist L, Nordell E, Jarnlo G-B, Stromqvist B, Wingstrand H, Thorngren K-G. Hip fractures in primary health care. *Scand J Prim Health Care* 1990;8:139-44.
5 Robbins JA, Donaldson LJ. Analysing stages of care in hospital stay for fractured neck of femur. *Lancet* 1984;ii:1028-9.
6 The old woman with a broken hip [editorial]. *Lancet* 1982;ii:419-20.
7 Mowat IG, Morgan RTT. Peterborough hospital at home scheme. *BMJ* 1982;284:641-3.
8 Parker MJ, Myles JW, Anand JK, Drewett R. Cost-benefit analysis of hip fracture treatment. *J Bone Joint Surg [Br]* 1992;74B:261-4.
9 Pryor GA, Myles JW, Williams DRR, Anand JK. Team management of the elderly patient with hip fracture. *Lancet* 1988;i:401-3.
10 Todd CJ, Williams DRR, Pryor GA, Parker MJ, Myles JW. Early discharge to "hospital at home" after fracture of neck of femur: psychosocial factors. In: Brenner G, Weber I, eds. *Health services research and primary health care.* Köln: Deutscher Ärzte-Verlag, 1991:185-90.
11 Pryor GA, Williams DRR. Rehabilitation after hip fractures. *J Bone Joint Surg [Br]* 1989;71B:471-4.
12 Parker MJ, Pryor GA, Myles JW. Early discharge after hip fracture. *Acta Orthop Scand* 1991;62:563-6.
13 Townsend P. *The last refuge.* London: Routledge and Kegan Paul, 1962.
14 Hodkinson HM. Evaluation of a mental test score for assessment of mental impairment in the elderly. *Age Ageing* 1972;1:233-8.
15 Drummond MF. *The principles of economic appraisal in health care.* Oxford: Oxford Medical Publications, 1980.
16 Russell EM. *Patient costing study.* Edinburgh: Scottish Home and Health Department, 1974. (Scottish Health Service Studies No 31.)
17 Reid N, Robinson G, Todd C. The quantity of nursing care on wards working 8-hour and 12-hour shifts. *Int J Nursing Studies* 1991;28:47-54.
18 Greatorex IF, Gibbs ACC. Proximal femoral fractures: some determinants of outcome. *J Epidemiol Community Health* 1988;42:365-9.
19 *British National Formulary.* No 22. London: British Medical Association, Royal Pharmaceutical Society of Great Britain, 1991.
20 Sommers LS, Schurman DJ, Jamison JQ, Woolson ST, Robison BL, Silverman JF. Clinical-directed hospital cost management for total hip arthroplasty patients. *Clin Orthop* 1990;258:168-75.
21 Office of Population and Census Surveys. *Hospital in-patient enquiry.* London: HMSO, 1987.

(Accepted 17 August 1993)

The hit and miss of ISS and TRISS

N Zoltie, F T de Dombal on behalf of the Yorkshire Trauma Audit Group

Abstract

***Objective*—To measure interobserver variation in recording injury from case notes and its effect on calculating injury severity scores (ISS) from identical data and predicting probabilities of survival by using the combined trauma and injury severity score (TRISS).**

***Design*—Observer variation study using injury severity scoring and subsequent calculation of probability of survival based on combined trauma and injury severity scores.**

***Subjects*—16 patients with a range of injury severity scores, and 15 observers.**

***Results*—There was a wide variation in recorded injury severity scores, the probability of two observers agreeing on the score being 0·28 (28%). The probability of any two observers agreeing over which severity band the patient should be in was 0·5 (50%). Observer variation was independent of the training and type of observer. Survival probability (calculated by combined trauma and injury severity scoring methodology from individual observers' scores) varied by over 0·2 in six of the 16 patients and by over 0·5 in three.**

***Conclusions*—There is wide observer variation in injury severity scoring, which highlights a potential fallibility in its use for trauma audit. The use of combined trauma and injury severity scoring for individual prediction of survival is potentially inaccurate except at the extremes of probabilities. The use of the 0·5 survival line on a combined trauma and injury severity score "pre-chart" is statistically and clinically inappropriate.**

Introduction

As a consequence of the recommendation of the Royal College of Surgeons Working Party,[1] increasing numbers of centres are conducting trauma audit. From time to time results are published for comparison and scrutiny.[2,3] The usual methodology used is the combined trauma and injury severity scoring system (TRISS),[4] which consists of calculations based on the injury severity score (ISS) and the revised trauma score (RTS). We report an observer variation study to establish the reliability and reproducibility of injury severity scoring and to ascertain what effect any variation might have on calculations of the probability of survival by means of combined trauma and injury severity scoring.

Patients and methods

Data from case notes of patients entered into the United Kingdom major trauma outcome study[5] from one hospital were used. As a completely unselected series might have resulted in a skewed distribution of injury severity scores measures were taken to ensure a wide spread of scores.

Patients began entering the United Kingdom major trauma outcome study on 1 April 1990, and data from the first 30 were screened. We selected the first four patients with low injury severity scores (0-20), as judged by the values actually entered into the United Kingdom major trauma outcome study; the first four patients with middle range scores (21-40); and the first four patients with high scores (41-75). Four other patients were selected at random so that observers would not know the exact numbers in each "group." Sixteen was the maximum number of cases that observers were thought able to code without time or fatigability problems. At that stage of selection neither other details (type of injury, area of injury, number of injuries) nor final outcome (death or survival) was known.

The case notes of the 16 patients were collected in their entirety and given to 15 observers for coding, no observer having knowledge of any other person's scores. Coding was carried out between January and July 1991. The observers were five accident and emergency consultants, six accident and emergency senior registrars, one accident and emergency registrar, and three trauma audit clerks. The observers were informed that there would be a range of severity and asked to identify every anatomical injury, code the injury (using the six figure code of abbreviated injury

Yorkshire Trauma Audit Group
Participants in the study are listed at the end.

Correspondence to:
Mr N Zoltie, Clinical Information Science Unit, Leeds University, 22 Hyde Terrace, Leeds LS2 9LN.

BMJ 1993;307:906-9

score (AIS90)), and score the injury and body area. No calculation of the injury severity score was made by the observers. The results were collected from all observers, the individual injury severity score calculated for each observer and each patient, and the results compared. Data for calculating the revised trauma score were also collected and the probability of survival calculated for every patient from each observer by means of combined trauma and injury severity scoring.

ANALYSIS OF DATA

Many statistical methods of measuring and assessing observer variation have been described,[6 7] and the subject is a matter of debate. The appendix describes some of the alternatives. For this study six specific practical questions were addressed:

(1) What is the probability of any one observer agreeing with another about the score of any individual patient? All individual scores were compared and the probabilities calculated of any two observers agreeing.

(2) What is the probability of any one observer placing the patient in the same banded group as a second observer? The standard audit classification suggested by the United Kingdom major trauma outcome study groups patients into seven bands based on injury severity scores. The results of each injury severity score calculation for each patient were grouped into the suggested bands and the probability of any two observers agreeing about banding calculated.

(3) Do observers agree over whether the patient has suffered major trauma? The accepted definition of major trauma is an injury severity score of 16 or over. The probability of any two observers agreeing on whether the patient had a score above or below 16 was calculated.

(4) Is variability in the injury severity score dependent on the type of patient or type of observer? Three groups of observers were compared: those who had attended a coding course but were not regularly coding; those who were actively coding regularly; and those who had not coded or attended a course but were aware of the methodology.

(5) Does the revised trauma score follow the same pattern? The procedures in *(1)* to *(4)* were repeated in respect of the revised trauma score.

(6) Does variability in the injury severity score affect calculations of probability of survival using the combined trauma and injury severity score? Each patient had probabilities of survival calculated on the basis of data recorded by each observer, and the results were compared. The probability of any two observers agreeing was calculated for each patient.

Results

INJURY SEVERITY SCORES

Figure 1 shows the spread of the injury severity scores calculated by each observer for each of the 16 patients. For various reasons four individual assessments were not completed, so that the figure shows a total of 236 scores for the 16 patients. There was considerable variation in the individual assessments and considerable difference in the spread of assessments. For example, there was almost unanimous agreement on scores in cases 1 and 10 but considerable variation between observers in cases 3, 7, 11, 14, and 16. In some cases the variation was extreme. In case 3, for example, two observers recorded the injury severity score as 10 and 17 whereas using identical data two others recorded it as 75. Almost identical values applied in case 11.

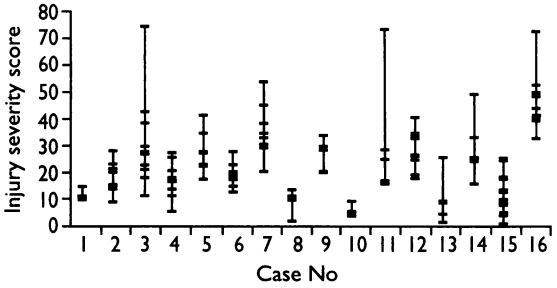

FIG 1—*Spread of injury severity scores for each patient*

In addition to recording actual injury severity scores, other studies have recommended "banding" these into several strata. Thus as well as summarising the raw data, figure 2 addresses each of the categorisations in turn. The top group shows the probability (for all 16 individual patients) of any two observers agreeing about the actual score (the answer to question 1).

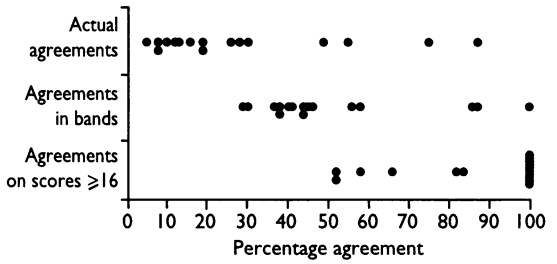

FIG 2—*Percentage agreement between observers over injury severity score*

The overall probability (for this series) was 28%, ranging from 90% in one patient to 10-20% in over half the patients. The second grouping shows the probability of any two observers agreeing about the severity band in which a patient should be placed (question 2), which was 51% overall and ranged from 100% agreement in one patient to 30-50% in 11 patients. The last grouping shows the probability of observers agreeing on whether the patient had major trauma (injury severity score ≥ 16). Though in most patients there was unanimous agreement, in six patients there was disagreement and in three of the 16 patients there was almost an even chance of assessment of the patient's trauma as major or minor (question 3).

In order to assess the effects of experience and education three different observer groups were compared (table). There was virtually no difference in agreement rates between accident and emergency consultants, senior registrars, and trauma audit clerks (question 4). Nor was there any significant difference between observers who had or had not attended a coding course. The agreement concerning the patients' "banding" was 47% overall among observers who had attended a coding course, 53% among those who had taught themselves, and 51% among those who were not actively coding nor had been on a course.

Percentage agreement between different groups of observers with respect to actual injury severity scores, banded injury severity scores, and injury severity scores of 16 or over

	Actual injury severity scores	Banded injury severity scores	Injury severity scores ≥16
Attended course, and coding actively (n=2)	25	60	81
Actively coding (n=4)	28	40	81
Attended course (n=4)	26	40	76
Not coding nor had attended course (n=5)	24	51	81
Significance	$\chi^2=0{\cdot}69$; $p>0{\cdot}5$	$\chi^2=6{\cdot}86$; $p>0{\cdot}05$	$\chi^2=7{\cdot}09$; $p>0{\cdot}05$

REVISED TRAUMA SCORES

Figure 3 shows the same analysis as in figure 1 but refers to revised trauma scores (same patients, same information source, same observers). In this case a completely different picture emerged. There was complete agreement among all observers with respect to revised trauma scores in nine of the 16 patients and

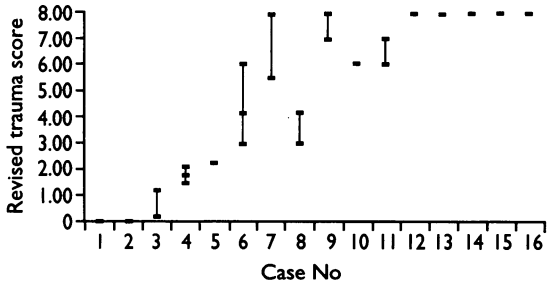

FIG 3—*Spread of revised trauma scores for each patient*

comparatively minor spread of calculated scores among the remainder (question 5).

EFFECT ON COMBINED TRAUMA AND INJURY SEVERITY SCORE

The combined trauma and injury severity score for survival probability is calculated by combining the revised trauma score and the injury severity score. Figure 4 shows the spread of survival probabilities in this series. In 11 of the 16 patients, despite variation in calculations of the injury severity score and revised trauma score, there was reasonable agreement about the probability of survival. In the other five patients, however, considerable variation was recorded between different observers' calculated values (question 6). Moreover, much of the significant variation occurred in the middle of the range—that is, between probabilities of survival of 0·1 and 0·9; fig 4).

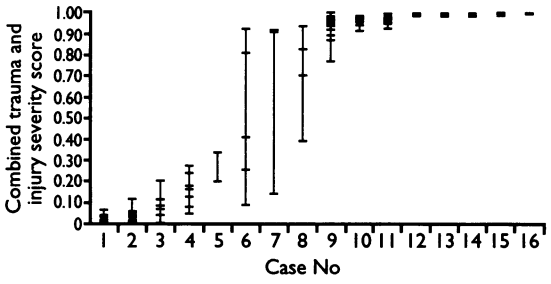

FIG 4—*Range of survival probabilities as estimated by combined trauma and injury severity score*

Discussion

The value of any scoring system used for assessment in clinical medicine relies heavily on the reliability and reproducibility of the method. The reliability of the combined trauma and injury severity score to predict survival has been well tested in North America, and the United Kingdom major trauma outcome study aims at testing the reliability of the North American database for use in the United Kingdom. The reproducibility of the method seems not to have been tested so widely previously.

Observer variation is inherent in most aspects of biomedicine,[7,8] though it is not the variation that matters but its extent and effect.[9,10] The findings of this study show the extent of the variation in injury severity scoring. Even when patients were grouped into bands of severity there remained a high probability of disagreement (around 50%) between any two observers on which band was appropriate, and this was unaffected by whether the observer had attended a coding course, was currently actively coding, or had done neither. When an even broader classification into two groups (above and below 16) was used there was still far from perfect agreement between observers on whether six of the 16 patients had or had not sustained major trauma.

The descriptive statistics used to demonstrate this variation do not distinguish the effect of one or two "bad" observers mixed with the majority of "good" observers. However, removal of any two observers' figures resulted in little reduction in variation, with no closer agreement of the remaining group. This is further exemplified by the division into groups of observers (table), where the means of each group did not differ either between groups or from the overall mean. Though not specified in the results, there was equal division of consultants, other doctors, and administrative assistants between the groups, and when these alternative groups were subanalysed there was no difference in mean agreements between these categories either. The variation in injury severity score thus seemed to be a genuine finding not attributable to specific characteristics of the observer.

Injury severity scoring is claimed to facilitate comparison of data from two or more different centres when it is used as a measure of severity case mix, which is essential for valid comparisons. Our finding that distinction in banding patients varied by up to 50% leads us to suggest that severity case mix assessed by the injury severity score may be subject to too much observer variation to be reliable. Clearly more detailed analysis is required to delineate the cause of observer variation. Preliminary analysis indicates that both omission and commission variation and interpretation are concerned and that the part of the body affected is comparatively unimportant. However, further detailed analysis will be undertaken.

Considerations regarding variability particularly apply when the injury severity score is used as a component of the combined trauma and injury severity scoring system for calculating the probability of survival.[4] We have shown the effect on the combined score of the variation in injury severity score. For those patients with probabilities of survival of between 0·05 and 0·95 there is a very large potential source of variation depending on the observer who collected the data. Combined trauma and injury severity score "precharts" have been suggested,[4] plotting the revised trauma score against the injury severity score with a line drawn along the 0·5 survival plots. Patients below this line are widely interpreted as survivors, and above as non-survivors, the exceptions often being audited as "unexpected survivals" and "unexpected deaths." Unfortunately, the widest variations or estimations in our series occurred precisely in this area. If our data are representative this use of a 0·5 survival line is probably widely inaccurate and statistically unsound.

It has been suggested that recording centrally by a single coder may overcome the variation in the injury severity score. This may or may not be the case for a single centre or location, but it might equally be a

Clinical implications

- The injury severity score is commonly used for audit of collected trauma cases
- Combined trauma and injury severity scoring methodology uses the injury severity score to calculate a probability of survival
- The injury severity score may be subject to considerable observer variation
- The combined trauma and injury severity score for probability of death also shows observer variation, which is less at the extremes of probabilities but potentially very large between p=0·05 and p=0·95.
- Comparisons between groups, hospitals, and countries must be treated with circumspection and great care and attention paid to collection of injury severity score data to reduce observer variation

source of major problems when comparisons are made between regions or countries.

We support the use of scoring systems in general terms and certainly do not wish to discourage the use of the injury severity score (or subsequently modified systems) for trauma audit. We suggest, however, that before any revision of the abbreviated injury score and the injury severity score observer variation should carefully be studied and measures adopted to minimise this problem for future users. Otherwise questions must remain over the precision of internal hospital audit, interhospital, and especially interregional or international, comparisons.

In conclusion, our study discloses some major problems with the methodology of the combined trauma and injury severity scoring system, but this does not undermine the need for national collection of trauma data by the United Kingdom major trauma outcome study. Rather, attention to the shortcomings we have identified should allow us—by placing the results in context—to utilise the results more appropriately and obtain the maximum value for these important national and international comparison.

We thank the participating members and in particular Mr J Sloan and Dr W Hulse for clinical input and help.

Participants in the study were: N Zoltie, M Clancy, J Hanson, and D Cartlidge (St James's Hospital, Leeds); M Williams, I Barlow, and J P Sloan (Leeds General Infirmary); A McGowan, A Gourdie, A Hawtin, and A K Marsden (Pinderfields Hospital, Wakefield); M Gibson and M Thirlway (York District Hospital); P Grout (Hull Royal Infirmary); and W Hulse (Harrogate District Hospital).

Appendix: statistical methods

Measures of assessing observer variation are the subject of much statistical debate. Fuller descriptions of many of the alternatives are provided elsewhere.[6,7] Most analyses in trials with multiple observers treat the resultant data as multiple two way comparisons, which was the method used for this study. For 15 observers this would result in 105 paired observations. If all the observers agreed, then the numbers of actual agreements would equal the total possible agreements —that is, the probability of agreement would be 1·0 (100%). For any number less than this the probability of agreement would equal the actual number of agreements divided by the total possible agreements, expressed as a percentage.

The most widely used coefficient of agreement in comparable studies is the κ statistic of Cohen.[11] There are, however, problems with the κ statistic. Cohen himself, for example, suggested several versions. As customarily used, κ measures the difference between observed agreement and the agreement that would be expected by chance in the same setting. A κ value of 0·4 generally represents reasonable agreement and 0·7 good agreement.

Our study discloses the limitations of this form of analysis. Firstly, the values recorded represented numerical assessments on non-numerical information. Secondly, the expected degree of agreement among the 15 observers varied greatly with the type of observation, ranging from the two way choice (major or minor trauma) to the actual score, for which the expected agreement was virtually nil. (In this setting the percentage of agreements was virtually equal to the κ statistic.) Finally, the κ statistic failed to take account of the clinical relevance of the data. Thus the data concerning agreement on major or minor trauma were impressive statistically but concealed an important factor—namely, that for six patients there was disagreement in up to half of comparisons on whether the patient had major or minor trauma and therefore on whether he or she should have been entered into a major trauma outcome study at all.

Brennan and Silman[7] have argued that for complex studies of observer variation more emphasis may have to be placed on raw data. We therefore present these in table A.

TABLE A—*Actual injury severity scores allocated to each patient by each observer*

Observer No	Case No*															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	9	14	29	17	34	17	38	13	29	4	29	25	4	16	25	45
2	9	13	29	17	22	14	45	10	29	4	25	34	9	25	8	50
3	9	13	29	4	27	22	38	10	29	4	16		9	50	18	54
4	9	20	10	16	27	12	45		29	4	75	34		50	4	75
5	9	13	38	25	22	27	38	9	29	4	29	26	9	34	19	51
6	13	22	29	20	22	19	29	13	34	4	25	18	9	50	9	34
7	9	14	42	26	34	14	20	10	20	4	25	27	26	50	26	54
8	9	19	17	10	17	12	30	1	20	4	17	34	1	26	10	50
9	9	13	75	25	27	17	33	9	20	4	29	41	9	25	13	41
10	9	13	38	17	41	18	33	9	20	4	25	26	9	25	5	50
11	13	13	26	13	17	12	29	13	29	4	75	41	9	50	18	42
12	9	27	75	16	22	22	34	10	29	9	29	19	9	50	10	50
13	9	13	27	16	22	17	54	13	29	4	75	35	4	50	14	50
14	9	8	20	16	17	22	34	10	29	4	16	18	9	16	10	50
15	9	13	22	16	27	17	54	10	29	4	25	34		50	9	75

*Graphs derived from table have been reordered, so that case numbers in table do not refer to those in derived graphs.

1 Commission on the Provision of Surgical Services. *Report of the working party on the management of patients with major injuries.* London: Royal College of Surgeons of England, 1988.
2 Phair IC, Barnes MR, Barton DJ, Allen MJ. Deaths following trauma: an audit of performance. *Ann R Coll Surg Engl* 1991;73:53-7.
3 Crawford R. Trauma audit: experience in north-east Scotland. *Br J Surg* 1991;78:1362-6.
4 Boyd CR, Tolson MA, Copes WS. Evaluating trauma care: the TRISS method. *J Trauma* 1987;27:370-7.
5 Yates DW, Woodford M, Hollis S. Preliminary analysis of the care of injured patients in 33 British hospitals: first report of the United Kingdom major trauma outcome study. *BMJ* 1992;305:737-40.
6 De Dombal FT, Softley A. IOIBD report No 1: observer variation in calculating indices of severity and activity in Crohn's disease. *Gut* 1987;28: 474-81.
7 Brennan P, Silman A. Statistical methods for assessing observer variability in clinical measures. *BMJ* 1992;304:1491-4.
8 Weiner N. Nonlinear prediction and dynamics. In: *Proceedings of 3rd Berkeley symposium on mathematical statistics and probability.* Vol 3. Berkeley: University of California, 1956:247.
9 Saiger GL. Observations on the probability of error in medical diagnosis. *Am J Intern Med* 1982;56:860-4.
10 Scheff TJ. Decision rules, types of error, and their consequences in medical diagnosis. *Behav Sci* 1963;8:97-107.
11 Cohen J. A coefficient of agreement for nominal scales. *Education and Psychological Measurement* 1960;20:37-46.

(Accepted 17 August 1993)

Correction

Short and long term prognosis of acute myocardial infarction since introduction of thrombolysis

An authors' error occurred in this paper by Robert Stevenson and colleagues (7 August, pp 349-52). In the results section of the abstract and the third paragraph of the subjects and methods section it is unclear how many patients were followed up after discharge from hospital. A total of 608 patients were studied, 89 died in hospital and 12 were lost to follow up after discharge. All 608 patients were followed up until hospital discharge or death in hospital and 507 were followed up after discharge from hospital.