Langenbeck's Arch Surg (2002) 387:55–58
DOI 10.1007/s00423-002-0283-0

E.A.M. Neugebauer
R. Lefering

Severity scores in surgery: what for and who needs them?

An introduction: definition, aims, classification and evaluation

Received: 3 August 2001
Accepted: 2 December 2001
Published online: 26 March 2002

This paper was presented at the 3rd Meeting of the Surgical Association for Clinical Research in Europe (SACRE), 16–18 November 2000, Norderstedt, Germany

E.A.M. Neugebauer (✉) · R. Lefering
Biochemical and Experimental Division,
II. Department of Surgery,
University of Cologne,
Ostmerheimerstrasse 200,
51109 Cologne, Germany
e-mail:
sekretariat-neugebauer@uni-koeln.de
Tel.: +49-221-989570
Fax: +49-221-9895730

Abstract Every patient represents a unique and complex situation a clinician has to deal with. In order to cope with this complexity of information, reduction is necessary, especially in communication about diseases or therapy. The first reduction is made when a patient is given a diagnosis which reflects a constellation of similar symptoms. A score also reduces the given amount of clinical data into a one-dimensional value. The primary aim of a score is a systematic comparison between patients and institutions. Scores reduce information to focus on the essentials. They are used for severity classification and prognosis, evaluation of outcome and treatment effects, case-mix adjustments in comparative audits, and economic evaluation. Quality criteria of score systems which should be considered in the development and application are: reliability, validity, measurability, applicability, and clinical relevance. This introductory article gives a brief description of these terms.

Keywords Score systems · Severity classification · Audit · Prognosis

Introduction

One of the consequences of high technology in medicine is an increasing amount of information such as laboratory results, function tests, physiological monitoring, and other derived data. To cope with this complexity and to use this information, appropriate abstraction seems to be inevitable. Scores, scales, and indices are attempts to reduce a complex clinical situation with its multiple dimensions into more compact data. The great advantage of this data reduction becomes obvious when regarding, for example, the very first score that was published in medicine, the Apgar score for newborn babies, or the ASA classification for surgical patients. Both give prognostic information and offer the possibility of comparisons between patients. Since then, hundreds of scores have been developed and published for different purposes, and the clinician is either confused which one to use, or even doubts the additional value of these scores. Surgeons often prefer their "gut feeling" than to rely on formal scores, e.g., for prediction of postoperative outcome immediately after operation [1]. Some argue that the calculation of scores takes too much extra time, and in retrospect it has rarely influenced clinical decision making.

A special symposium at the 3rd Meeting of the Surgical Association for Clinical Research in Europe (SACRE) held on 16–18 November 2000 in Norderstedt, Germany, was therefore organized to discuss the question: "Severity scores in surgery – what for and who needs them?" This introductory article is intended to give some basic information about scoring systems, their aims, classification, development, and evaluation.

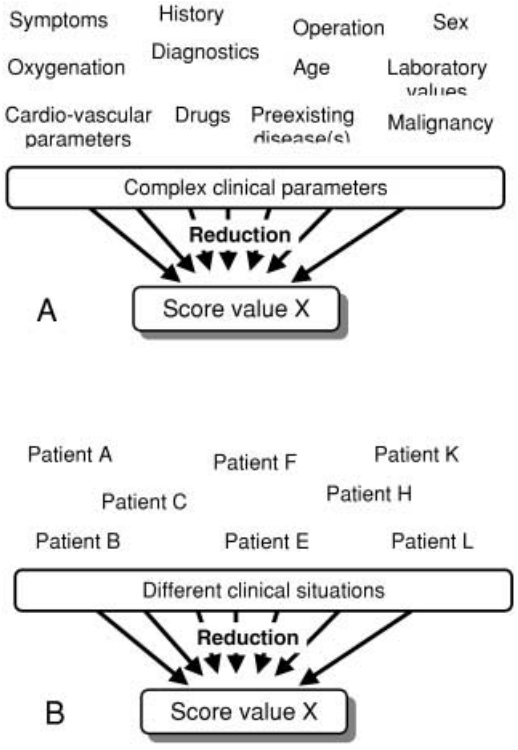

Fig. 1 **A** A score is an attempt to reduce a complex clinical situation with multiple dimensions into a one-dimensional value. **B** As a consequence, a single score value represents different clinical situations

Terminology

A scale, like a thermometer, is an instrument to measure a clinical phenomenon; a score is a value on the scale in a given patient. Clinical scales provide a standardized, repeatable measurement of a patient's condition or functional status, just as a thermometer provides a standardized, reproducible measurement of temperature [2].

This definition is not uniformly accepted throughout the literature – synonyms of a scale are a scoring system, or classification or staging system. However, the intention is the same all over: it is an attempt to reduce a complex clinical situation with its multiple dimensions into a one-dimensional value (Fig. 1). As a sequence, different clinical situations are given the same score value. This value usually reflects a ranking in a specific system, e.g., risk of death or complication, which allows for comparisons between patient groups. Simple scales contain only one question or element and are designed to measure only one phenomenon. More complicated scales consist of a number of separate elements or questions, that cover different issues [2]. In addition, most prognostic scores include some basic data like age and comorbidity to improve prediction. The response to each of the elements are assigned weighted point values which are then summed up to get a total score for each patient [2, 3].

Table 1 Reducing information to the essentials

Application of scores
– Severity classification
– Prognosis and outcome prediction
– Evaluation of treatment
– Quality assessment and audit
– Economic evaluation
Quality criteria
– Reliability
– Validity
– Measurability
– Applicability
– Clinical relevance

Severity scores in surgery. What for?

Every reduction means loss of information. On the other hand, a surplus of information may mask the relevant knowledge. One of the advantages of a score is reduction of information to the essentials. A score enables one to gain objective and reproducible information across different patients at definite points of time. This approach can be used for different purposes (Table 1).

Severity classification and prognosis

Classification of the severity of a disease means the description of groups by risk. It is of confirmed value in controlled and uncontrolled clinical trials [4]. Scores can be used to generate a prognosis because the risk of a negative outcome increases with increasing severity of disease. Whether and how scores which were mainly developed for risk stratification can be used for the prediction in individual cases is still highly controversial. In our opinion, score-based risk of death estimates are not justified in the individual patient [3, 5].

Treatment and outcome evaluation

Evaluative scores (scales) measure the change or stability in a population over time, particularly the effect of a therapeutic intervention [2]. To evaluate a therapeutic effect the patient's pre-operative scale rank can be compared with the postoperative value. Using sequential assessment of outcome, comparisons could use the maximum score value in a defined period of time (e.g., Apache II score [6]), the cumulative sum of score values (e.g., sum of TISS-28 points over time [7]), or the duration of organ failure. Quality of life instruments [8] and

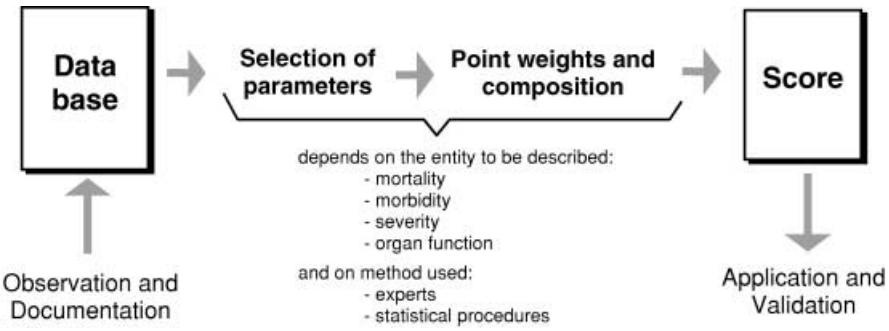

Fig. 2 Steps in the development of a score system

pain measurement scales [9] are further examples for effective use of score systems for outcome evaluation in clinical trials.

Audit and quality assurance programs

Scoring systems play an important role in hospital audit programs since internal and external comparisons are useless without any information about the severity of disease. The increasing popularity of hospital league tables (bench marking) make scoring systems a mandatory tool for case mix description and adaptation. Comparisons of ICU performance often use the standardized mortality ratio (SMR) which relates the observed mortality to an expected rate based on score predictions [10]. In multiple-injury patients, the TRISS score is widely used for national and international comparison [11, 12].

Economic evaluation

Some scores are useful for cost and cost-effectiveness studies. For example, the TISS-28 score [13] tries to quantify the use of intensive care resources with a point system. The EUROQOL [14] is a quality of life measurement instrument which generates a one-dimensional index useful for cost-effectiveness analyses. But application of scores in this field are still rather limited.

Classification and development of scores

Scores can be used for different purposes. When selecting an appropriate scoring system, the user first has to decide what the score primarily should do: stratification of severity; prediction of outcome; quantification of resource use; or assessment of outcome. Scores can further be classified according to different criteria:

a) The degree of *generalizability*, which means a general score is applicable across different diseases while a specific score mainly considers the characteristics of one selected disease or organ. Well known general scores are Apache II (for intensive care patients) and the ASA classification (for surgical patients). The more specific a score is the more detailed the information about the patient will be, at the cost of a decreased applicability.
b) The *composition* of a score, especially, which components are used to generate the score value. Some scores only consider physiological parameters, while other also include therapeutic activities as indirect indicators of an altered physiology.
c) The *type of application*, i.e., whether the score was developed and validated for a single assessment (e.g., pre- or postoperatively, or on admission to ICU), or whether the primary intention was a sequential monitoring and a detection of daily changes.
d) The *method of development*, which can either be subjectively based on experts' opinion or group consensus, or more empirical based on a large number of real observations, analyzed by statistical procedures like logistic regression analysis (Fig. 2). However, the development of a score is always based on prior experience, either implicit (experts) or explicit (database). The question of which approach is superior is controversial, but a combination of both might be the best solution.

Evaluation of scores

Independent of its development or its intended application, a score must fulfill several prerequisites before it is used in clinical practice, scientific trials, or audit programs. In order to find out whether a score is appropriate for a specific situation, a number of aspects have to be considered. These aspects include quality characteristics of the measurement instruments, as well as its applicability. Criteria for the assessment and selection of scores are detailed below, as well as in Table 1 [3].

Reliability

Reliability describes the exactness with which a score measurement can be performed, i.e., how accurate a

score can be reproduced. Reliability is a formal criterion, focussing on *how* something is measured, and not what the measurement means. A score can be very reliable but without any importance. Criteria for good reliability include: well-defined items, clear and unequivocal choice of measured values, and defined instructions in case of missing values. To prove the reliability of a measurement tool, a test-retest examination can be performed, or the application of the score by different persons can be compared (inter-observer variation).

Validity

A score is valid if it actually does measure what it intends to measure. One proof for validity is the so-called face validity, which means that the used parameters 'obviously' correspond with the aim of the score. A more formal method to check the validity of severity scores is to build subgroups of increasing score values, which should, on average, show a consistent increase in morbidity and mortality. On the other hand, patients with different clinical outcome (e.g., survivor and non-survivor) should demonstrate respective differences in initial severity score values.

Measurability

The measurability of a score is determined by the availability of the parameters and the time required to calculate the score. The more sophisticated parameters a score includes and special laboratory test or calculations are required, the more the measurability may decrease.

Applicability

Applicability compares whether the examined disease (or patient group) of interest is comparable to the situation the score initially was designed for. If the patients are different, validity checks have to precede an application of the score. If application of a score in a different patient population has not been reported in the literature before, an own validation study has to be performed.

Clinical relevance

When choosing a score, one also has to consider the interpretation of its results. In clinical trials, differences in score values have to be clinically relevant. A score value is only useful and understandable to someone who works intensively with this score. Therefore, a generally accepted and frequently applied score should be preferred to a newly developed one.

References

1. Hartley MN, Sagar PM (1994) The surgeon's "gut feeling" as a predictor of post-operative outcome. Ann R Coll Surg Engl 76 [Suppl 6]:277–278
2. Carlson ME, Johanson NA, Williams PG (1991) Scaling, scoring and staging. In: Troidl H, McKneally MF, Mulder DS, Wechsler AS, McPeek B, Spitzer WO (eds) Surgical research – basic principles and practice. Springer, Berlin Heidelberg New York, pp 192–200
3. Neugebauer E, Lefering R (2000) Scores. In: Burchardi H, Larsen R, Schuster HP (eds) Intensivmedizin. Springer, Berlin Heidelberg New York, pp 83–94
4. Ohmann C, Wittmann DH, Wacha H and the Peritonitis Study Group (1993) Prospective evaluation of prognostic scoring systems in peritonitis. Eur J Surg 159:267–74
5. Suter P, Armaganidis A, Beaufils F, Beaufils F, Bonfill X, Burchardi H, Cook D, Fagot-Largeault, Thijs L, Vesconi S, Williams A (1994) Predicting outcome in ICU patients. Consensus conference organized by the ESICM and the SRLF. Intensive Care Med 20:390–397
6. Pilz G, Fateh-Moghadam S, Viell B, Bujdoso O, Döring G, Marget W, Neumann R, Werdan K (1993) Supplemental immunoglobulin therapy in sepsis and septic shock – comparison of mortality under treatment with polyvalent i.v. immunoglobulin versus placebo. Theor Surg 8:61–83
7. Lefering R (1999) Biostatistical aspects of outcome evaluation using TISS-28. Eur J Surg 163 [Suppl 584]:56–61
8. Eypasch E, Wood-Dauphinée S, Williams JI, Ure B, Neugebauer E, Troidl H (1993) Der gastrointestinale Lebensqualitätsindex (GLQI). Ein klinimetrischer Index zur Befindlichkeitsmessung in der gastroenterologischen Chirurgie. Chirurg 64:264–274
9. Hebebrand D, Troidl H, Spangenberger W, Neugebauer E, Schwalm T, Gunther MW (1994) Laparoskopische oder klassische Appendektomie? Eine prospektiv randomisierte Studie. Chirurg 65:112–120
10. Grover FL, Hammermeister KE, Shroyer ALW (1995) Quality initiatives and the power of the database: what they are and how they run. Ann Thorac Surg 60:1514–1521
11. Champion HR, Copes WS, Sacco WJ, Lawnick MM, Keast SL, Bain LW, Flanagan ME, Frey CF (1990) The major trauma outcome study: Establishing national norms for trauma care. J Trauma 30:1356–1365
12. Bouillon B, Neugebauer E, Rixen D, Lefering R, Tiling T (1996) Wertigkeit klinischer Scoresysteme zur Beurteilung der Verletzungsschwere und als Instrument für ein Qualitätsmanagement bei Schwerverletzten. Zentralbl Chir 121:914–923
13. Reis Miranda D (1999) Outcome assessment – TISS as a tool to evaluate cost-effectiveness of immunological treatment. Eur J Surg 163 [Suppl 584]:51–55
14. Kind P (1996) The EuroQoL instrument: an index of health-related quality of life. In: Spilker B (ed) Quality of life and pharmacoeconomics in clinical trials. Lippincott-Raven, Philadelphia, pp 191–202