HOME

MedicalBiostatistics.com

Medical Scoring

Method of Scoring

Characteristics that have gradient are relatively easy to score than the measurements on nominal scale. If the gradient is already numeric such as pain on visual analogue scale, the score is immediately obtained but the problem arises for ordinal factors that have gradient but no numeric scale is available. Examples are severity of disease, which is categorized as mild, moderate, serious, critical, and degree of satisfaction categorized as completely dissatisfied to fully satisfied. Characteristics on nominal scale such as presence or absence of signs and symptoms in any case defy quantitation at individual level. The method of assigning numerics to ordinal and nominal characteristics should be such that can reduce uncertainties and not introduce additional epistemic uncertainties.

Method of Scoring for Graded Characteristics

Most simple scoring is linear such as 0 for no disease, 1 for mild disease, 2 for moderate disease, 3 for serious disease and 4 for critical condition. Such scoring assumes that the difference between mild and no disease is the same as between critical and serious disease. It is legitimate to ask for such scoring that why scores 0, 3, 6, 9 and 12 are not more appropriate, which also are linear, or even why geometric scoring such as 0, 1, 2, 4, 8 is not better. Very few studies have been carried out to investigate various alternatives and thus nothing can be stated with confidence. Nevertheless, 0, 1, 2, 3, 4 type remains the most widely used scores because of their simplicity. They go unnoticed. No explanation is generally required when such simple scores are used. Any other scores are expected to be accompanied by justification. A more complex procedure to generate scores is illustrated in Example 1.

Example 1 Scoring for motor components in multiple sclerosis

Mobility suffers in multiple sclerosis and various scales are used to assess functionality in the patients. Among those relatively easy to do are time taken to perform the 10-m timed walk (TMTW) and nine-hole peg test (NHPT) for right and left hand. Vaney et al. [1] developed short and graphic ability score (SaGAS) as (2×TMTW + NHPTright + NHPTleft) after taking logarithm of these timed values. The authors demonstrated good correlation of SaGAS with established tests such as Multiple Sclerosis Functional Composite, Expanded Disability Status Scale and Rivermead Mobility Index. Features of SaGAS such as simple, intuitive, and nonphysician-based measure were enumerated while pleading its used in multiple sclerosis patients.

Side note: Functionality in multiple sclerosis patients has gradient and its metric measurements has always been a challenge. The authors developed an easy to implement scoring system. Its good correlation with other more complex scoring methods is one way to demonstrate its validity. For a real validity assessment, comparison with a gold standard is more convincing. Such gold may not exist in this case; in most cases the gold might be too cumbersome and expensive. Equivalence of the new scoring with the gold when demonstrated would establish its real validity.

Method of Scoring for Diagnosis

Scoring might help in situations where diagnosis or differential diagnosis is difficult and requires a lot of expertise that may not be immediately available. See example on hypothyroid diagnosis that uses scores on clinical signs and symptoms. In that example, the scoring helps to establish or rule out hypothyroidism in nearly half the cases. Thus the need of further investigations is reduced to one half. For another example, see Guo et al. [2] that provides scoring system for diagnosis of Hirschsprung's disease in the neonatal period. Rosen et al. [3] proposed a smaller 5-item index as a diagnostic tool for erectile dysfunction compared with the larger 15-item scoring. This uses linear scores for each item.

Signs-symptoms are qualitative and they play a significant role in establishing diagnosis. Converting such qualities to a numerical score has always been a challenge and no widely acceptable method is available yet. The following examples and discussion are based on the methods used by some workers who ventured into this area.

Example 2 Delphi method of scoring

Consider measuring nursing workload in an ICU by a scoring system. Thus this is not for diagnosis. Yamase [4] assessed this workload by 88 items relating to (a) number of nurses required, (b) muscular exertion, (c) mental stress, (d) skill, and (e) intensity. A three-round Delphi survey among 20 skilled ICU nurses assigned consensus four-grade (0 to 3) score to each of the 73 items after excluding 15 items considered unnecessary. These scores were confirmed by surveying 118 nurses in other ICUs. The 'comprehensive nursing intervention score' is the sum of all these individual item scores. The scoring system was confirmed as reflecting true workload by applying it to the daily care of 107 patients.

Example 2 illustrates how a simple method can be used to develop a scoring system. Validation of the individual scores by another group of nurses and of the scoring system by using it in actual conditions tends to enhance the confidence in the scoring system.

There are examples of assigning arbitrary scores. Thurnau et al. [5] assigned 'weighted numerical scores of increasing magnitude' on the basis of degree of abnormality for early identification and assessment of severity of pregnancy-induced hypertension. The scoring system was found to correspond well with the clinical status.

A common and acceptable method of assigning scores is based on the regression coefficients that are estimated using the multiple regression when the outcome is quantitative. For qualitative outcomes, particularly dichotomous, such as disease present/absent, logistic regression coefficients are used. The factors surmised to determine the outcome are antecedents whose coefficients in the logistic regression equations are significant. You would soon see that they can be interpreted as log of odds ratio (OR). Larger the odds ratio, better is the predictive utility of the factor. Thus the score can be assigned in proportion of the OR to those factors that turn out to be significant predictors. The method is illustrated in Example 3. This also is more on gradation of disease than on diagnosis although the outcome is dichotomous, death or survival.

Example 3 Scoring system to stratify risk in unstable angina

Unstable angina is a complex syndrome prognosticated by a host of factors such as age, hypertension, diabetes, hypercholesterolemia, smoking, previous myocardial infarction, ST segment deviation, troponin test, etc. Piombo et al. [6] studied a large number of such factors in 473 patients and found four of them significant predictor in a multivariate logistic regression of in-hospital occurrence of refractory ischemia, acute MI, or death. These three together formed unfavorable outcome. The ORs were 4.03, 2.29, 2.21 and 2.0 for ST segment deviation, age ≥ 70 years, previous coronary artery bypass grafting, and positive troponin test (T ≥ 0.1 mg/ml), respectively. The scores assigned were 4, 2, 2 and 2 corresponding to the respective ORs. The highest possible score was 10. It was divided into three categories: 0 or 2 for low risk, 4 or 6 for intermediate risk, and 8 or 10 for high risk. Under this scoring system, a score of 1, 3, 5, 7, or 9 is not possible.

The scoring system was validated in another group of 242 patients that provided similar results. Nearly 63% of patients were assigned to low risk group, 31% to intermediary risk and 6% to high risk group. Predictive power of the scoring system assessed by C-statistic (area under the ROC curve) was 0.72. The C-statistic is one of the important criteria that determine the validity of the scoring system.

Side note: The authors called score categories 0 to 2, 4 to 6, and 8 to 10 as tertiles, which is not a correct use of the term since these categories do not comprise one-third subjects each that tertiles would. Also the predictive power of 72% as measured by the area under the ROC curve is not adequate to inspire confidence. The authors have discussed limitations of their study such as patients were selected for a trial with strict inclusion and exclusion criteria that may have compromised representativeness. In practical application, many patients of unstable angina may not have previous angiography performed and this variable would not be available. In addition, serum cardiac markers were not well defined.

Example 3 was chosen not because it provides a valid scoring system but because it uses an appropriate method for selection of factors and for assigning them proper score. There are many other examples of this type. Purasiri et al. [7] combined the results of clinical examination, mammogram, ultrasonograph, and fine needle aspiration cytology by assigning them weighted scores using stepwise logistic regression. This combined score performed better than any of them individually in differentiating malignancy from benign lesion in suspected cases of breast cancer. In this study, the confirmed diagnosis was later available so that the 'gold' was present. However, the authors used the term index and not score. Another example is the scoring system developed by Chiu et al. [8] for early detection of oral submucous fibrosis based on self-administered questionnaire. This had C-statistic 0.90. Rassi et al. [9] also assigned scores proportional to the regression coefficients to the independent significant factors for death in Chagas heart disease. The C-statistic was 0.84.

The method of assigning scores proportional to the regression coefficients is valid only when the values of the predictive factors are standardized to mean zero and variance one before the regression is run. As demonstrated later in the chapter on regression, without this standardization, the regression coefficients are not comparable and are severely affected by the unit of measurement.

Since regression coefficient after standardization just stated measures the contribution of the factor to the outcome, this method of scoring looks at least face-valid as it assigns higher score to the factor that contributes more to the outcome. Also the regression is able to identify the factors that are significant independent contributors and need to be included in the scoring system. Thus this method has some desirable properties. However, it may fail to provide a valid scoring system in some situations as discussed next.

Validity and Reliability of a Scoring System

Although qualities such as easy to understand and easy to implement can be cited, basic statistical qualities of scoring systems are validity and reliability. Of these two, validity is more important and difficult to assess too. If a scoring system were not valid, its good reliability would seldom be useful.

Validity of Scoring System

How to assess that a scoring system is providing the right result? First, it should look just about right. It should also correspond well to the knowledge of experts. If a scoring system surprises you and the experts, reconsider the elements that are causing this surprise. But the most important assessment of validity is against the gold standard. Basic difficulty in this assessment is in identifying and implementing the gold standard against which the validity is checked. If the gold is easy to perform, there is no need of a scoring system. In the case of pregnancy-induced hypertension, Thurnau et al. [5] compared the scoring results with clinical manifestation. *If clinical manifestation is to be considered gold and if clinical assessment is relatively easy, nothing additional is gained by the scoring system.* Scoring system is useful only when it really adds to the clinical picture, or when it replaces a complicated procedure. The latter can happen when final diagnosis is based on consensus of experts or when it emerges later in the course of disease. An explicit advantage is the objectivisation scores introduce that may lack in clinical assessment.

When the results from gold standard are not available, the worth of a scoring system is assessed using alternatives. One is to see whether scoring system gives results that are consistent with undisputable outcomes such as death. Rassi et al. [9] reported for their scoring system for predicting death in Chagas heart disease that patients with low (0 to 6 points) score had 10 percent 10-year mortality rate, with medium (7 to 11 point) score had 44 percent, and with high (12 to 20 points) score had 84 percent. This provides an indirect evidence of validity of the scoring. Note in this case that the gold in this case is death and there is no way to assess it in advance except by prognostic factors summarized by total score.

Second aspect of validity is establishing it in a different sample. This in fact testifies repeatability. If another sample of similar nature gives similar results, it is safe to conclude that the scoring system is not sample-specific and has at least some generalizability. In almost all examples discussed in this section, the **validation sample** is different from **development sample,** and the results were shown to replicate adequately.

Third type of validation of a scoring system is its comparison with an established and more cumbersome system. This is to check if an easier version provides the same results. Both could be excellent or both could be poor but that is not the issue in this kind of comparison. Thus the comparison is not with a gold standard. This concurrent validity provides evidence that the easier version can replace the cumbersome procedure. Moreno and Morais [10] compared 28-question simplified therapeutic intervention scoring system (TISS) with the standard 72- question TISS for nursing workload in intensive care units, and came up with the conclusion that the simplified version is just as good. Evans et al. [11] developed scoring system for identifying BRCA½ mutation and found that it outperforms existing models.

The statistical performance of a scoring system is generally judged by the C-statistic that measures area under the ROC curve. Recall that an area of 0.5 indicates the scoring system is not helpful in properly identifying the disease or any health condition. An unattainable area of 1.0 implies a perfect system. In between, area between 0.70 and 0.79 is considered as satisfactory, between 0.80 and 0.89 as good, and an area 0.90 or more as excellent. Very few scoring systems, for that matter any diagnostic aid, will attain an area 0.90 or more.

Reliability of a Scoring System

In addition to being valid, any medical assessment tool should also be reliable in the senses of repeatability and reproducibility. Reliability may suffer if the wordings are not precise and instructions are not explicit so that there is a room for subjective interpretation—either by the assessor or by the assessee, or by both.

The reliability of a scoring system is assessed interobserver (also called interrater) as well as intraobserver. For both these assessments, intraclass correlation is used. This correlation must be in excess of 0.90 for good reliability and a value between 0.80 and 0.89 is acceptable. Any tool with less than 0.80 is suspect. Use scoring system with such low intraclass correlation with caution.

Example 4 Reliability of scores for severity of rickets

Severity of nutritional rickets can be assessed by the degree of metaphyseal fraying and cupping, and the proportion of the growth plate affected, based on radiographs of wrists and knees. Thacher et al. [12] evaluated the utility and reproducibility of a 10-point scoring system that progresses in half-point increments from zero (normal) to 10 points (most severe). They found that interobserver correlation of the score was 0.84 or greater for all observer pairs used by them, and intraobserver correlation was 0.89 or greater for each observer. Thus there is a fair amount of consistency. The authors conclude that the score should be useful to objectively assess the severity of rickets.

REFERENCES

1. Vaney C, Vaney S, Wade DT. SaGAS, the Short and Graphic Ability Score: An alternative scoring method for the motor components of the Multiple Sclerosis Functional Composite. *Mult Scler* 2004; 10:231-242.
2. Guo W, Zhang Q, Chen Y, Hou D. Diagnostic scoring system of Hirschsprug's disease in the neonatal period. *Asian J Surg* 2006; 29:176-179.
3. Rosen RC, Cappelleri JC, Smith MD, et al. Development and evaluation of an abridged 5-item version of the International Index for Erectile Function (IIEF-5) as a diagnostic tool for erectile dysfunction. *Int J Imp Res* 1999; 11:319-326.
4. Yamase H. Development of a comprehensive scoring system to measure multifaceted nursing workloads in ICU. *Nurs Health Sci* 2003; 5:299-308.
5. Thurnau GR, Dyer A, Depp OR 3rd, Martin AO. The development of a profile scoring system for early identification and severity assessment of pregnancy-induced hypertension. *Am J Obstet Gynecol* 1983; 146:406-416.
6. Piombo AC, Gagliardi JA, Guelta J, et al. A new scoring system to stratify risk in unstable angina. *BMC Cardiovasc Disord* 2003; 3:8.
7. Purasiri P, Abdalla M, Heys SD, et al. A novel diagnostic index for use in the breast clinic. *JR Coll Surg Edinb* 1996; 41:30-34.

8. Chiu CJ, Lee WC, Chiang CP, Hahn LJ, Kuo YS, Chen CJ. A scoring system for the early detection of oral submucous fibrosis based on a self-administered questionnaire. *J Public Health Dent* 2002; 62:28-31.

9. Rassi A Jr, Rassi A, Little, WC, et al. Development and validation of a risk score for predicting death in Chagas heart disease. *N Engl J Med* 2006; 355:799-808.

10. Moreno R, Morais P. Validation of the simplified therapeutic intervention scoring system on an independent database. *Intensive Care Med* 1997; 23:640-644.

11. Evans DG, Eccles DM, Rahman N, et al. A new scoring system for the chances of identifying a BRCA1/2 mutation outperforms existing models including BRCAPRO. *J Med Genet* 2004; 41:474-480.

12. Thacher TD, Fischer PR, Pettifor JM, Lawson JO, Manaster BJ, Reading JC. Radiographic scoring method for the assessment of the severity of nutritional rickets. *J Trop Pediatr* 2000; 46:132-139.

Adapted from Medical Biostatistics, Fourth Edition
by A. Indrayan (http://indrayan.weebly.com)
Chapman & Hall/CRC Press, 2018
(https://www.routledge.com/Medical-Biostatistics/Indrayan-Malhotra/p/book/9781498799539)

Details of the book and reviews available at http://MedicalBiostatistics.synthasite.com