Intensive Care Med (2001) 27: 1365–1369
DOI 10.1007/s001340101012

ORIGINAL

Kees H. Polderman
Edward M. F. Jorna
Armand R. J. Girbes

Inter-observer variability in APACHE II scoring: effect of strict guidelines and training

Received: 2 June 2000
Final revision received: 2 June 2000
Accepted: 18 December 2000
Published online: 14 July 2001

K. H. Polderman (✉) · A. R. J. Girbes
Surgical Intensive Care Unit,
University Hospital Vrije Universiteit,
P. O. Box 7057, 1007 MB Amsterdam,
The Netherlands
E-mail: k.polderman@azvu.nl
Phone: +31-20-4443912
Fax: +31-20-6392125

E. M. F. Jorna
Department of Anesthesiology,
University Hospital Vrije Universiteit,
Amsterdam, The Netherlands

Abstract *Objective:* To assess the effect of strict guidelines and a rigorous training program on variability in scoring the revised Acute Physiology and Chronic Health Evaluation (APACHE II).
Design and setting: Prospective survey and intervention in the surgical ICU of a university teaching hospital.
Measurements: Seven experienced intensivists and nine residents determined APACHE II scores in one set of patients before and in another set 4 months after a rigorous training program, following strict guidelines for using the APACHE II.
Results: APACHE II scores were 14.3 ± 4.4 before the training program (n = 12) and 18.9 ± 2.4 after (n = 11). Interobserver agreement rates increased significantly from 59.7 % to 76.5 % and the interobserver reliability coefficient (weighted κ) from 0.72 to 0.85 after our training program was implemented. The changes were significantly greater in experienced intensivists than in less experienced residents, indicating that more experienced physicians profited to a greater degree from our training program.
Conclusion: Interobserver variability in APACHE II scoring decreases markedly when strict guidelines and a regular training program are implemented, particularly among more experienced physicians. However, in our study a degree of variability (10–15 %) persisted even in experienced intensivists with similar training, experience, and background, suggesting that a degree of variability is inherent in APACHE II scoring.

Keywords Acute Physiology and Chronic Health Evaluation II score · Interobserver variability · Sources of errors · Training program · Guidelines · Intensive care unit performance

Introduction

The revised Acute Physiology and Chronic Health Evaluation (APACHE II) score [1] is used widely in intensive care units (ICUs), as an index of case severity, to predict outcome, to assess clinical performance and quality of care in the ICU [2, 3], and in research protocols to ascertain differences between treatment groups. In spite of its general acceptance and widespread use, the variability and validity of APACHE II scoring have not been extensively studied. To our knowledge no studies have dealt specifically with the effect of training programs on reliability of APACHE II scoring. We have previously reported wide interobserver variability in APACHE II scores among a group of physicians assessing the same patients [4]. The study included experienced intensivists, who according to the literature, can be regarded as experts in the use of scoring systems, and ICU residents. No significant difference in interobserver variability was found between the two groups of

physicians [4]. Subsequently we analyzed our data and identified various causes of error and confusion in APACHE II scoring. We then set about to correct these problems by implementing a rigorous training program and imposing strict guidelines in the use of the APACHE II. We report here the effect of this training program and the implementation of strict guidelines on variability in APACHE II scoring.

The main objectives of our study were: (a) to identify common sources of errors in APACHE II scoring and to devise appropriate guidelines and training program to overcome them; (b) to evaluate the effect of our training program on interobserver variability; and (c) to determine the degree of variability that persists in spite of training (probably indicating inherent variability in the APACHE II score). Additional objectives were to compare interobserver variability between residents and experienced intensivists and to assess whether the change in interobserver variability differs between experienced intensivists and residents.

Methods and materials

Sixteen physicians (seven experienced intensivists and nine ICU residents with a mean ICU experience of 6 months) determined APACHE II scores from the charts of 12 suitable patients before the training program and those of 11 thereafter. This produced 16 APACHE II scores of each individual patient. The patients were randomly selected; cardiac surgery patients were excluded because the APACHE II score has not been validated for this category of patients. Patients with ICU stay of 8 h or less were also excluded.

Medical records in our surgical ICU were kept manually by residents, nurses, and intensivists. All physicians and nurses were instructed to keep extensive records in a uniform way; grand rounds were carried out every day, which included inspection of the charts by the staff intensivist on duty. Mean arterial pressure and heart rate were measured continuously, and the most abnormal value of the past hour was noted at least once every hour. Alarms were set to go off if blood pressure and/or heart rate exceeded or fell below predetermined limits; these abnormal values were then noted while appropriate corrective action was taken. Thus the most abnormal value of the past hour was always recorded. Core temperature was measured continuously, using a rectal probe. Extensive laboratory tests were performed every 4 h on a routine basis. Blood gasses, hematocrit, and electrolytes were assessed every hour, or more often if necessary, using a blood gas analysis machine (Chiron 855, Chiron Diagnostics, Houten, The Netherlands) located in the ICU. All laboratory test results were noted in the chart by the nurse. In addition, printed laboratory results were included in the patient chart twice every day. Residents and nurses work in 8-h shifts, and a complete report of the patients' condition was noted during each shift. All decisions made at the grand rounds, changes in the patients' condition, test results, and other relevant issues were noted in the chart.

Before the study APACHE II scores were assessed by intensivists and residents on the basis of a manual on scoring issues in the ICU. This manual contained, among other things, the original APACHE II paper [1] and selected other articles from international and national medical journals dealing with scoring issues. A copy of this manual was located at the desk in the ICU where physicians wrote their APACHE II assessments. Various ICU reference books were available in our ICU library, which is located on the same floor as the ICU itself. In addition, there were educational sessions twice yearly for ICU residents and fellows dealing with the issue of quality control in the ICU including scoring issues (although much less thoroughly and specifically than in our later training sessions).

Each batch of scores was collected over a 5-week period. Following the first part of our study (assessment of variability before implementation of our training program) we carefully studied sources of problems and confusion that had contributed to high variability in APACHE II scoring. Based on this analysis we drafted strict guidelines and implemented a rigorous training program for our medical staff. The new guidelines were clarified during these training sessions, and various examples were provided and discussed. In addition, a set of "difficult" patients (actual patients admitted to the ICU and "'hypothetical" patients) were scored during training sessions, and the results and pitfalls discussed extensively. All physicians were instructed to adhere strictly to the original guidelines as laid down by Knaus et al. [1]. For example, points *must* be given for a single measurement of high blood pressure or for a brief episode of tachycardia, even when these data are inconsistent with the general trend, and regardless of whether the physician feels that the measurement accurately reflects the patient's physiological status. Similarly, data acquired in the emergency room or operating room before ICU admission *must* be excluded. Strict guidelines were laid down regarding chronic health points and points for Glasgow Coma Score. We developed a consensus as to how all physicians would define acute renal failure leading to doubling of "renal points." In addition, we provided all physicians with written guidelines and a quick reference table for calculating points from the APACHE II oxygenation formula. During training sessions, difficult "hypothetical" cases were discussed, and issues over which confusion was apparent were incorporated in the written guidelines. In addition, an abbreviated version of these guidelines was printed on the back of the forms used for APACHE II scoring. After a few months we again asked our medical staff to score a number of consecutive patients, to assess the effect of our training program on interobserver variability. The same methods and period for score collection were used.

Statistical analysis used Students' unpaired t test for comparison of standard deviations before and after training. Excel and SSPS 9 software for Windows was used for further statistical analysis including 95 % confidence interval (CI) and univariate analysis to determine the weighted κ scores and interobserver agreement rates before and after training.

Results

Table 1 presents the findings before the implementation of guidelines and training program (group 1; $n = 12$) and Table 2 the findings after the implementation (group 2; $n = 11$). Overall interobserver variability in APACHE scores decreased significantly, as demonstrated by the decrease in standard deviation from 4.4 to 2.4, with means of 14.3 and 18.9, respectively. Before our training program there was no difference in scoring variability between inexperienced residents and experienced intensivists, but thereafter the variability was significantly lower in experienced intensivists. Detailed results for individual patients are shown in Table 2. The 95 % CI val-

Table 1 Variability in APACHE II scoring before training program

	All physicians (n = 16)				Residents (n = 9)				Experts (n = 7)			
Patient no.	Mean	95% CI	Median	Range	Mean	95% CI	Median	Range	Mean	95% CI	Median	Range
1	15.4	11.9–18.9	15	7–21	15.9	11.4–20.4	16	7–21	14.8	11.5–18.1	15	9–17
2	18.2	14.1–22.3	18	7–24	17.9	13.9–21.9	16.5	7–22	18.6	14.5–22.7	18	8–24
3	8.6	6.3–10.9	9	3–19	8.2	5.9–10.6	7.5	4–19	9.1	6.9–22.7	10	3–14
4	14.0	10.8–17.2	14.5	8–19	13.6	10.6–16.6	13	8–17	14.5	11.1–17.9	15	9–19
5	13.8	11.4–16.3	14	7–19	13.0	10.4–15.6	13	8–19	14.8	12.5–17.1	15	7–18
6	12.4	10.3–14.5	12	8–15	12.8	10.5–15.2	13.5	9–15	12.2	10.2–14.2	11.5	8–13
7	15.2	12.1–18.3	13.5	5–24	16.1	12.7–19.5	15	5–24	14.0	11.2–16.8	13	6–19
8	10.8	9.2–12.4	12	4–15	10.9	9.1–12.3	10	4–14	10.1	9.3–12.5	9	7–15
9	21.4	17.2–25.6	10	8–34	22.1	17.7–26.5	22	8–34	20.5	16.6–24.4	18.5	13–34
10	19.1	15.3–22.9	19	8–30	18.2	14.4–22.0	18	8–30	20.3	16.4–24.2	20	11–26
11	12.3	10.0–14.7	11.5	6–20	12.6	10.3–15.0	12	6–20	11.9	9.5–14.3	12	6–18
12	10.4	8.1–12.8	10	2–16	9.8	7.3–12.3	10	2–16	11.2	9.1–13.3	12.5	5–15
All	14.3 ± 4.4	10.8–18.2[a]	13.9	–	14.3 ± 4.4*	10.9–17.9[a]	13.3	–	14.3 ± 4.3*	10.1–18.6[a]	14.3	–

* NS residents vs. experts
[a] Mean deviation from the median

Table 2 Variability in APACHE II scoring after implementation of training program

	All physicians (n = 16)				Residents (n = 9)				Experts (n = 7)			
Patient no.	Mean	95% CI	Median	Range	Mean	95% CI	Median	Range	Mean	95% CI	Median	Range
1	23.0	22.0–24.0	24	19–25	23.8	22.4–25.1	24.5	19–25	22.3	20.9–23.7	22	20–25
2	34.1	31.9–36.2	34.5	25–40	32.8	22.5–36.0	33	33–40	35.3	32.5–38.1	36	29–40
3	19.4	18.0–20.9	18	16–25	19.6	17.6–21.7	18.5	16–25	19.5	17.2–21.3	18	19–24
4	15.8	14.3–17.3	15	14–27	16.3	13.7–19.0	15	14–27	15.1	14.1–16.1	15	14–18
5	18.2	17.1–19.3	18	15–23	18.2	16.4–20.1	19	15–23	18.1	16.9–19.3	18	15–20
6	16.1	15.0–17.2	16.5	13–19	16.3	14.6–17.9	16	13–19	16.0	14.5–17.5	16.5	12–19
7	16.2	14.9–17.4	16	12–21	16.9	15.0–18.8	17	14–21	15.4	13.9–17.0	16	12–18
8	15.2	14.3–16.1	15	12–20	15.8	14.2–17.3	15	12–20	14.6	13.7–15.5	15	13–16
9	14.5	13.7–15.3	15	11–16	14.7	13.6–15.8	15	11–16	14.4	13.1–15.6	15	11–16
10	26.1	25.3–26.9	26	24–31	26.4	25.1–27.6	26	25–31	25.9	24.8–27.0	26	24–29
11	9.1	8.1–9.9	9	6–12	10.1	9.1–11.1	11	7–21	7.9	6.7–9.0	7.5	6–11
All	18.9 ± 2.4	14.5–26.1	16.2	–	19.2 ± 2.8*	14.7–26.4	16.9	–	17.3 ± 2.0*	14.4–25.9	16.0	–

*$p < 0.02$ residents vs. experts

ues decreased considerably in both residents and intensivists, indicating an increased reliability of the score. The decrease in 95% CI values was greater in intensivists.

To test the overall agreement in the physicians' observations we also calculated weighted κ scores and interobserver agreement rates (i.e., how many scores attributed by physicians were exactly the same, expressed as a percentage of the total number of scores). These values are shown in Table 3. Before implementation of our training program the overall interobserver agreement was 59.7% (residents, 59.2%, experts, 60.1%; NS). The overall κ score was 0.72 (intensivists 0.74, residents 0.71). After implementation of our training program the overall interobserver agreement increased to 76.5% (residents 70.9%, experts 76.5%; $p < 0.01$ residents vs. experts, $p < 0.01$ before vs. after training overall and in both groups). The overall κ score increased to 0.85 (intensivists 0.89, residents 0.82).

The most frequent sources of problems in APACHE scoring before implementation of our training program were:

- Inclusion of data acquired in the operating room and/or emergency room (some physicians, both experts and residents, mistakenly took these data into account). On the other hand, data obtained in the operating room in patients requiring surgery within the first 24 h after ICU admission, which should have been included in APACHE score assessment, were often mistakenly disregarded.
- Interpretation of data which were inconsistent with the general trend (for example, tachycardia which was found only once during a 24-h period was (mis-

Table 3 Reliability coefficients and interobserver agreement rates for APACHE II scoring before and after implementation of training program

	Before training program			After training program		
	Agreement	κ	95 % CI	Agreement	κ	95 % CI
Residents	59.2 %	0.71	0.64–0.78	70.9 %*,**,***	0.82	0.74–0.9
Intensivists	60.1 %	0.74	0.66–0.80	79.8 %*,**,***	0.89	0.81–0.95
Overall	59.7 %	0.72	0.68–0.76	76.5 %*	0.85	0.80–0.90

* $p < 0.01$ before vs. after training program, ** $p < 0.01$ residents vs. experts, *** $p < 0.01$ difference in training effect (= increases in interobserver agreement) experts vs. residents

takenly) disregarded by some physicians but not by others.
- The attributing of chronic health points (2 or 5); this was the most frequent source of error. Many physicians erroneously attributed points for chronic diseases not listed in the original APACHE II score, such as stable angina of New York Heart Association class II, mild emphysema, diabetes mellitus, and any malignancy.
- The definition of what constitutes "emergency surgery."
- Erroneous exclusion of some laboratory results (in particular those determined directly in the ICU using a rapid laboratory device installed in the ICU (Rapidlab 855, Chiron Diagnostics, Emeryville, Calif., USA) instead of by the general hospital laboratory; these laboratory results were mistakenly disregarded by many physicians.
- Glasgow Coma Score: mistaken attribution of points for loss of consciousness induced by sedation, or mistaken nonattribution of points in patients with loss of consciousness induced by medication taken in suicide attempt.

Less frequently occurring errors were:
- Pacemaker present upon admission; points for bradycardia incorrectly attributed.
- Missing data scored as normal.
- High serum creatinine before admission: points mistakenly doubled.
- Nonattribution of points for abnormal values of hematocrit, white blood cell count, creatinine or sodium because abnormal laboratory values were found even prior to ICU admission, or because levels remained in the same range during the first 24 h after ICU admission.
- Counting errors in addition of APACHE II score and errors in attribution of points for age.
- Calculating errors in oxygenation formula (alveolar-arterial O_2 gradient) in patients with FIO_2 greater than 50 %.

Persisting sources of variability in scoring after implementation of our training program included:
- Attribution of chronic health points (differences in interpretation of low-dose prednisone treatment, human immunodeficiency virus status, recent surgical intervention for colon cancer).
- Artificial cooling in a patient with severe head injury.
- Incomplete or missing data (two patients).
- Mistaken exclusion of data inconsistent with the general trend (in spite of our training program!).
- Mistaken exclusion of data obtained in the operating room in a patient requiring surgery in the first 24 h after ICU admission.
- Calculation errors in computing APACHE scores.

Conclusion

We conclude that interobserver variability in APACHE II scoring decreases markedly when strict guidelines and a training program are implemented. A significant increase in interobserver agreement was observed (from 59.7 % to 76.5 %); weighted κ scores increased from 0.72 to 0.85. Expressed as a percentage of APACHE scores, variability decreased from 31 % to 13 %. In theory, decreases in variability could be explained if the second group of patients were in some way "easier" to score than the first group; however, both groups of patients were randomly selected, and similar degrees of variability (between 10 % and 20 %) were observed in all patients after training. This makes it highly likely that the observed effects were due to our training program. Experienced physicians profited more from this program than inexperienced residents. However, a significant degree of variability in scoring persisted, even in experienced intensivists with similar training, experience, and background. Therefore we conclude that some degree of variability is inherent in APACHE II scoring. We found this variability to be 10–15 %; the figure was somewhat lower in experienced intensivists.

Previous studies dealing with APACHE II application and implementation have provided some data on variability [5, 6, 7, 8]. However, only few of these studies dealt specifically with score variability and reliability, focusing instead on the appropriateness of the

APACHE II equation from the United States for other countries [8], reliability and reproducibility in the reporting of clinical parameters [9], quality and reliability of APACHE II data collection in regard to the reliability of mortality prediction by the score [7], and comparison in interobserver variability between physicians and nurses [6]. To our knowledge, no previous studies have dealt with the effects of a training program and guideline implementation on scoring variability in everyday clinical practice. Moreover, only few data are available on problems encountered in specific items of the APACHE II score, although some authors have touched upon this subject [7, 9]. We chose not to perform statistical subanalysis on the sources of variability listed in our results in view of the relatively small number of patients included in our study and because this was not the primary aim of our investigation.

In some studies, usually those carried out in the United States, scoring was performed out by experts specialized in full-time data collection and assessment of APACHE scores. Such experts would obviously have fewer or no benefits from a training program and guideline implementation such as described in this study. However, this is not the normal situation in most hospitals, where APACHE II scores are assessed by the attending physician and/or by the supervising consultant. Therefore we feel that our findings more adequately reflect the situation in most ICUs.

In view of its key role in ICU medical literature, it is remarkable that methods sections in papers reporting APACHE scores rarely if ever describe how and by whom these scores were obtained, or whether the scores were revised or checked by the authors. An additional problem with the APACHE II score is that it is likely to lead to serious bias against ICUs in hospitals with well-equipped emergency rooms. The reason for this is that stabilization of very ill patients in the emergency room results in lower APACHE scores at ICU admission. In contrast, hospitals in which such patients are admitted directly to the ICU are likely to show a better "performance" based on APACHE II scores. This potential bias can further complicate comparisons in performance between different ICUs, and the interpretation of APACHE II scores in medical literature.

Although the APACHE II score is an extremely valuable tool in critical care medicine, it is important to realize that a degree of variability is inherent to the score (as is generally the case in research tools). Physicians treating patients in the ICU, those interpreting the results of clinical trials, and financial controllers and decision makers should be aware of these limitations in the interpretation and use of APACHE II scores, and the inherent variability of the score. Firm assessments of quality of medical care based (largely) on scoring systems should be viewed with some caution. Potential authors of ICU outcome studies should describe how and by whom scores were assessed, and take score variability into account when analyzing their results. This applies especially to multicenter ICU outcome trials, where slight differences in scoring procedures between the participating centers may increase variability in scoring.

Acknowledgements We thank Dr. E. De Lange-De Klerk, medical statistician, for her help in data analysis and statistical evaluation, and Dr. J.M. Dixon for his critical appraisal of the manuscript.

References

1. Knaus WA, Draper E, Wagner DP, Zimmerman JE (1985) APACHE II: a severity of disease classification system. Crit Care Med 13: 818–829
2. Lemeshow S, Le Gall JR (1994) Modelling the severity of illness of ICU patients. JAMA 272: 1049–1055
3. Suter P, Armaganidis A, Beaufils F, et al (1994) Predicting outcome in ICU patients: consensus conference organised by the ESICM and the SRLF. Intensive Care Med 20: 390–397
4. Polderman KH, Thijs LG, Girbes ARJ (1999) Intra-observer variability in the use of the APACHE II scoring system. Lancet 353: 380
5. Féry Lemonnier E, Landais P, Loirat P, Kleinknecht D, Brivet F (1995) Evaluation of severity scoring systems in ICUs – translation, conversion and definition ambiguities as a source of inter-observer variability in Apache II, SAPS and OSF. Intensive Care Med 21: 356–360
6. Holt AW, Bury LK, Bersten AD, Skowronski GA, Vedig AE (1992) Prospective evaluation of residents and nurses as severity score data collectors. Crit Care Med 20: 1688–1691
7. Chen LM, Martin CM, Morrison TL, Sibbald WJ (1999) Interobserver variability in data collection of the APACHE II score in teaching and community hospitals. Crit Care Med 27: 1999–2004
8. Rowan KM, Kerr JH, Major E, McPherson K, Short A, Vessey MP (1993) Intensive Care Society's APACHE II study in Britain and Ireland. II. Outcome comparisons of intensive care units after adjustment for case mix by the American APACHE II method. BMJ 307: 977–981
9. Damiano AM, Bergner M, Draper EA, Knaus WA, Wagner DP (1992) Reliability of a measure of severity of illness: acute physiology of chronic health evaluation. II. J Clin Epidemiol 45: 93–101