INCREASING RELIABILITY OF APACHE II SCORES IN A MEDICAL-SURGICAL INTENSIVE CARE UNIT: A QUALITY IMPROVEMENT STUDY

By Laura Donahoe, BHSc, MD, Ellen McDonald, RN, Michelle E. Kho, BHSc(PT), MSc, Margaret Maclennan, RN, Paul W. Stratford, MSc, and Deborah J. Cook, MD, MSc, FRCPC

**Background** Given their clinical, research, and administrative purposes, scores on the Acute Physiology and Chronic Health Evaluation (APACHE) II should be reliable, whether calculated by health care personnel or a clinical information system.

**Objective** To determine reliability of APACHE II scores calculated by a clinical information system and by health care personnel before and after a multifaceted quality improvement intervention.

**Methods** APACHE II scores of 37 consecutive patients admitted to a closed, 15-bed, university-affiliated intensive care unit were collected by a research coordinator, a database clerk, and a clinical information system. After a quality improvement intervention focused on health care personnel and the clinical information system, the same methods were used to collect data on 32 consecutive patients. The research coordinator and the clerk did not know each other’s scores or the information system’s score. The data analyst did not know the source of the scores until analysis was complete.

**Results** APACHE II scores obtained by the clerk and the research coordinator were highly reliable (intraclass correlation coefficient, 0.88 before vs 0.80 after intervention; \( P = .25 \)). No significant changes were detected after the intervention; however, compared with scores of the research coordinator, the overall reliability of APACHE II scores calculated by the clinical information system improved (intraclass correlation coefficient, 0.24 before intervention vs 0.91 after intervention, \( P < .001 \)).

**Conclusions** After completion of a quality improvement intervention, health care personnel and a computerized clinical information system calculated sufficiently reliable APACHE II scores for clinical, research, and administrative purposes. (American Journal of Critical Care. 2009;18:58-64)
In a previous study,9 we documented that base-
line APACHE II scores collected by 2 research clerks
and an ICU research coordinator had excellent reli-
ability (intraclass correlation coefficient [ICC], 0.90).
However, 2 APACHE II components, the CHI, and
the verbal component (GCS-V) of the GCS score were
less reliable (ICC, 0.65 and 0.40, respectively). Pol-
derman et al7 improved the reliability of APACHE II
scores from 0.71 to 0.85 through standardized data
collection and specific training sessions. Using prin-
ciples similar to those applied by Polderman et al,7
we sought to improve the less reliable components
of the APACHE II score in our ICU.

This prospective before-and-after study had 3
objectives: (1) document the reliability of APACHE
II scores recorded by a clinical information system,
a database clerk, and a research coordinator, (2)
implement a multifaceted, multidisciplinary quality
improvement intervention to improve the reliability
of APACHE II scores, and (3) reevaluate the reliabil-
ity of APACHE II scores after the intervention.

Materials and Methods

This study was conducted in a university-affiliated,
15-bed, medical-surgical ICU at St Joseph's Healthcare
in Hamilton, Ontario. In this setting, APACHE II
scores are calculated automatically by the bedside clinical
information system for clinical
purposes and by other personnel for
research and administrative purposes.

Baseline Data Collection (2 Months)

We previously reported the data
collection methods for the baseline
phase of the study.9 Briefly, we
recorded APACHE II scores calcu-
lated for consecutive patients admit-
ted to the ICU by a database clerk
and a research coordinator. We
excluded patients if their ICU stay was less than 24
hours. In addition, we collected APACHE II scores
from our bedside clinical information system, Care-
Vue Classic (CareVue, Philips, Andover, Massachu-
setts), which provides new baseline data for this
report. CareVue is an electronic medical record sys-
tem for critically ill patients that collects data on
vital signs, ventilation settings, intravenous infu-
sions, nursing and medical assessments, and labora-
tory values. Data are uploaded hourly unless otherwise
specified by bedside nurses. APACHE II data ele-
ments were set to autocalculate daily.

Quality Improvement Intervention (5 Months)

The focus of the intervention was improving
data collected by health care personnel and the clinical
information system. We divided the intervention
The intervention focused on improving data collected by health care personnel and the clinical information system.

Interventions targeted at chronic health and GCS items did not result in significant changes.

Scores only for patients enrolled in clinical studies. For the purposes of this study, while the intervention was occurring, we did not analyze any APACHE II scores.

Reevaluation (3 Months)

The data collection methods during the reevaluation phase were the same as those used before the intervention. A different data clerk, who was blinded to the APACHE II score calculations and source, entered information from the 3 different raters into a database. The database clerk and research coordinator had no knowledge of each other’s scores or of the CareVue scores before and after the intervention. Because human performance can improve when people are aware that their behavior is being observed (the Hawthorne effect) or evaluated (the sentinel effect), both before and after the intervention, the bedside nurses were unaware of the conduct of the study. However, because the purpose of the intervention was to improve APACHE II documentation, we explicitly exposed the bedside nurses to the 6 components of the quality improvement intervention.

Patient care was at the discretion of the ICU team throughout the study. This study was approved by the St. Joseph’s Health Care Research Ethics Board, which waived the need for informed consent because the study did not affect patient care.

Sample Size Calculation and Data Analysis

We calculated interrater reliability by using the ICC, and we calculated ICCs for the APACHE II (total score, APS, age, and CHI) and GCS score (total, verbal, motor, and eyes) components. For each phase, we calculated a sample size of 32 patients to test whether an obtained reliability of 0.90 exceeded a reliability of 0.80, given 3 raters, a 1-tailed α = .05, and a power of 80%.

To ensure we had sufficient observations, we enrolled an additional 5 patients. Reliability was classified as follows: slight, 0.0-0.20; fair, 0.21-0.40; moderate, 0.41-0.60; substantial, 0.61-0.80; and almost perfect, 0.81-1.00.

We compared ICCs between each pair of raters before and after the intervention. We explored differences in ICC from before to after the intervention by using the Bonferroni correction (for 10 comparisons, our critical P value was .005). All tests were 2 sided. We calculated 95% confidence intervals where appropriate.

We calculated descriptive statistics and used t tests and Wilcoxon rank-sum tests to compare continuous data and χ² tests to compare proportions. We used SPSS (version 14, SPSS Inc, Chicago, Illinois) for all analyses. The data analyst had no knowledge of the source of the scores until analyses were complete.
### Table 1
Overview of quality improvement interventions to improve reliability of scores on the Acute Physiology and Chronic Health Evaluation (APACHE) II

<table>
<thead>
<tr>
<th>Intervention Description</th>
<th>Redesign Issue</th>
<th>Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. <strong>CareVue reconfiguration</strong></td>
<td>Chronic Health Index (CHI) visibility: The CHI questions were not available for staff to answer in CareVue.</td>
<td>We placed the CHI questions on the initial default screen to minimize the systematic underestimation of CHI noted before this intervention.</td>
</tr>
<tr>
<td></td>
<td>Timing of APACHE II calculation: APACHE II scores were calculated automatically at 7:30 AM. Results of the 6 AM blood tests were not always available for inclusion, and the calculation ran even if a patient had been admitted to the intensive care unit for less than 24 hours.</td>
<td>We reset the timing to run 24 hours after a patient’s admission to the intensive care unit to ensure consistent collection of patients’ information.</td>
</tr>
<tr>
<td></td>
<td>Modification of calculation variables</td>
<td>Solution</td>
</tr>
<tr>
<td></td>
<td>Temperature: CareVue recognized only core temperatures in degrees Fahrenheit.</td>
<td>First, we added a calculation to record the rectal or core temperature in degrees Celsius. Second, we added 0.5°C if the temperature was oral and 1.0°C if it was axillary or tympanic.</td>
</tr>
<tr>
<td></td>
<td>Blood pressure readings: The APACHE II scoring did not consider noninvasive measurements of blood pressure.</td>
<td>We incorporated an automatically interfaced noninvasive measurement of blood pressure if an arterial catheter was not in place.</td>
</tr>
<tr>
<td></td>
<td>Respiratory rate: The APACHE II scoring did not include respiratory rates of patients receiving mechanical ventilation.</td>
<td>We added a calculation to include respiratory rate for patients receiving mechanical ventilation if they had no spontaneous respirations.</td>
</tr>
<tr>
<td></td>
<td>Laboratory units: The CareVue system originated in the United States and calculated laboratory results on the basis of conventional (US) units; however, we collected our data in SI (Systeme Internationale) units.</td>
<td>We converted APACHE II laboratory parameters from SI units to conventional US units for creatinine and hematocrit laboratory values.</td>
</tr>
<tr>
<td></td>
<td>Acute renal failure: The calculation of APACHE II scores was incorrect for patients in acute renal failure who had abnormally high creatinine levels on laboratory tests.</td>
<td>We inserted the correct calculation to double the APACHE II points for a high creatinine level in the presence of acute renal failure (eg, a serum level of creatinine &gt;305 mmol/L would score 4 points, but in the presence of acute renal failure it would score 8 points).</td>
</tr>
<tr>
<td></td>
<td>Oxygenation: The APACHE II scoring considered only PaO2 oxygenation values and did not calculate values for the alveolar-arterial oxygen difference if the fraction of inspired oxygen was &gt;50%.</td>
<td>We modified the calculation so that if a patient’s fraction of inspired oxygen is &gt;50%, then values for the alveolar-arterial oxygen difference are calculated for inclusion in the APACHE II score, whereas if the fraction of inspired oxygen &lt;50%, PaO2 is included in the APACHE II score.</td>
</tr>
<tr>
<td></td>
<td>pH: The APACHE II scoring only considered arterial pH and not serum bicarbonate values in the absence of arterial pH.</td>
<td>We modified the calculation to use the serum bicarbonate value if an arterial pH was not available for inclusion in the APACHE II score.</td>
</tr>
</tbody>
</table>

2. **Education**
   - Documentation in-service training sessions: Our nurse educator and nurse informatician conducted training sessions focused on documenting the CHI components and the verbal component of the score on the Glasgow Coma Scale in intubated patients.
   - Educational materials: We posted information sheets and electronic resources at each computer workstation.

3. **Point-of-care electronic reminders**
   - For all new admissions, twice daily we sent computer-generated electronic reminders to each patient’s computer workstation to complete the CHI.

4. **Local opinion leaders, champions**
   - Local nurse opinion leaders and charge nurse champions prompted bedside nurses to complete CHI documentation for all new admissions.

5. **Audit and feedback**
   - The nurse informatician provided regular, informal audit and feedback to bedside nurses, reinforcing the need for timely documentation on all data for the APACHE II score.

6. **Policy**
   - The unit’s working group approved and disseminated, through the hospital’s intranet, a formal policy written by the CareVue Quality Team regarding timely completion of the CHI within 24 hours of a patient’s admission to the intensive care unit.
Results

We enrolled 37 patients before the intervention and 32 patients after the intervention. We detected no significant differences in patients’ characteristics from before to after the intervention (Table 2). Both before and after the intervention, the reliability of the APACHE II scores generated by the database clerk and the research coordinator remained almost perfect, with no significant change over time. However, we did detect initial deficiencies in the reliability of the APACHE II scores generated by the CareVue system. Before the intervention, the reliability of the CareVue APACHE II scores was fair compared with the reliability of the scores generated by the database clerk (ICC [95% confidence interval], 0.29 [0.0, 0.61]) and the research coordinator (0.24 [0.0, 0.59]); however, the reliability between the scores from the database clerk and the research coordinator was classified as almost perfect at 0.88 (0.77, 0.94).

After the intervention, the reliability of the CareVue APACHE II scores compared with the scores from the research coordinator was almost perfect and significantly improved at 0.91 (0.82, 0.95). Compared with the scores of the database clerk, the CareVue APACHE II scores also improved in reliability ($P = .03$; Table 3).

When the database clerk and the research coordinator were compared, we did not detect any significant improvements in reliability of either the CHI or the GCS-V subscales of the APACHE II scores after our multifaceted interventions (Table 3). The reliability of the CHI before the intervention was 0.65 (0.42, 0.80), whereas the reliability after the intervention was 0.35 (0.01, 0.62). The reliability of the GCS-V score before the intervention was 0.44 (0.11, 0.67), which improved to 0.59 (0.31, 0.77), although this difference was not significant.

The remaining major subscales of the APACHE II score, age, APS, and GCS, had no significant changes in reliability between the database clerk and the research coordinator from before to after the intervention (Table 3). Age scores were almost perfect, and although the APS and GCS reliability scores were somewhat lower after the intervention, this difference was not significant. Compared with before the intervention, the Eye subcomponent of the GCS score had significantly worse reliability after the intervention at 0.51 (0.20, 0.73); however, the reliability was still moderate. Following data collection, we examined the distribution of the CHI and each of the GCS components across patients and found little variability in scores, a situation that might decrease the ability to detect change over time.

Discussion

After a multifaceted, multidisciplinary intervention, we detected significant improvement in the reliability of APACHE II scores calculated by a clinical information system. We also found that total APACHE II scores obtained by a database clerk and a research coordinator are highly reliable and consistent over time. Likewise, all of the major components of the APACHE II score except the CHI were reasonably reliable and consistent over time. After our specific interventions targeted at improving the CHI component and GCS-V subscales, however, we did not detect significant changes.

Numerous strategies to change behavior have been suggested to improve the quality of health care.7 We selected interventions that were most likely to address the problems we observed, building on previous work on behavior change in our ICU as well as the published literature on practice improvement.
as summarized in several systematic reviews and adapted to our limited budget and setting. Our quality improvement intervention focused primarily on bedside nurses, because they directly influence patient care and are heavily involved in documenting patients’ illness. Key components of successful quality improvement projects are leaders and champions.14 In our study, leadership was provided by a nurse informatician, nurse manager, and nurse educator; champions were charge nurses who encouraged and modeled accurate and timely documentation of the CHI component of the APACHE II score, which required input from bedside nurses. Our multifaceted approach included educational meetings and materials, point-of-care electronic reminders, local opinion leaders, prompts, auditing, and feedback. The goals of the project were encoded in a formal unit policy that was posted and endorsed by the multidisciplinary CareVue quality team and ICU working group.

Our study has limitations. In any multifaceted intervention of this type, it is difficult to determine which component was responsible for the greatest change in behavior. We hypothesize that the reconfiguration of CareVue had the greatest impact, and of the other components, we think that the reminders, prompts from peer leaders, auditing, and feedback had the most important role in increasing the completion of the CHI questions. Certainly, changing the CareVue system calculations through automation was important. In this study we did not use a clinical decision support system, a powerful method of changing behavior,15 because we were not using an information system to support clinical decision making for patient care. Neither the CHI nor the GCS-V subcomponents of the APACHE II score were designed to discriminate among patients; thus, documenting significant improvements in the reliability of these subcomponents may not be possible because of the minimal variation across patients. Because we used a computerized clinical information system, our results are not applicable to paper-based bedside records of measures of illness severity, and the reliability of APACHE II scores would most likely be lower among newly hired personnel. Finally, although our results are generalizable to similar medical-surgical ICUs with a wide variety of admission diagnoses, they may not necessarily be generalizable to exclusively neurosurgery, cardiac surgery, or trauma ICUs.

Strengths of our project include the consistent team that participated in all 3 phases of the research program. We involved professionals from many disciplines, including staff from nursing informatics, management, physicians, and research personnel, thereby ensuring that we incorporated diverse suggestions from a broad range of perspectives in designing our intervention. We minimized selection bias by enrolling consecutive patients who met entry criteria before and after the intervention. We conducted this study prospectively, thus avoiding errors and incomplete records associated with retrospective chart review. No data were missing. We used blinded data abstraction, entry, and analysis. The implementation strategies we used ranged from simple to complex and were readily available, well accepted, and

### Table 3

<table>
<thead>
<tr>
<th></th>
<th>Research coordinator vs database clerk</th>
<th>Research coordinator vs CareVue</th>
<th>Database clerk vs CareVue</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Before</td>
<td>After</td>
<td>P</td>
</tr>
<tr>
<td><strong>Total APACHE II score</strong></td>
<td>0.88 (0.77, 0.94)</td>
<td>0.80 (0.62, 0.90)</td>
<td>0.25</td>
</tr>
<tr>
<td>Age</td>
<td>0.97 (0.94, 0.98)</td>
<td>1.00</td>
<td>NA (division by 0 error)</td>
</tr>
<tr>
<td>CHI</td>
<td>0.65 (0.42, 0.80)</td>
<td>0.35 (0.01, 0.62)</td>
<td>0.10</td>
</tr>
<tr>
<td>APS</td>
<td>0.89 (0.77, 0.94)</td>
<td>0.76 (0.50, 0.88)</td>
<td>0.08</td>
</tr>
<tr>
<td>Total GCS</td>
<td>0.83 (0.67, 0.91)</td>
<td>0.61 (0.33, 0.79)</td>
<td>0.05</td>
</tr>
<tr>
<td>Verbal</td>
<td>0.44 (0.11, 0.67)</td>
<td>0.59 (0.31, 0.77)</td>
<td>0.40</td>
</tr>
<tr>
<td>Motor</td>
<td>0.86 (0.75, 0.93)</td>
<td>0.61 (0.33, 0.79)</td>
<td>0.02</td>
</tr>
<tr>
<td>Eyes</td>
<td>0.86 (0.74, 0.92)</td>
<td>0.51 (0.20, 0.73)</td>
<td>.003&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

**Abbreviations:** APACHE, Acute Physiology and Chronic Health Evaluation; APS, Acute Physiology Score; GCS, score on Glasgow Coma Scale; NA, not applicable.

<sup>a</sup> Values in Before and After columns are intraclass correlation coefficient (95% confidence interval).

<sup>b</sup> After adjustment for multiple comparisons, significantly different from before to after intervention (2-tailed).

<sup>c</sup> APACHE II components (eg, Age, CHI, APS) were not collected from the CareVue system in phase 1.

**Computerized charting systems are integral to health care institutions and must perform reliably.**
eLetters

Now that you’ve read the article, create or contribute to an online discussion on this topic. Visit www.ajcconline.org and click “Respond to This Article” in either the full-text or PDF view of the article.

REFERENCES

To purchase electronic or print reprints, contact The InnoVision Group, 101 Columbia, Aliso Viejo, CA 92656. Phone, (800) 809-2273 or (949) 362-2050 (ext 532); fax, (949) 362-2048; e-mail, reprints@aacn.org.
Increasing Reliability of APACHE II Scores in a Medical-Surgical Intensive Care Unit: A Quality Improvement Study
Laura Donahoe, Ellen McDonald, Michelle E. Kho, Margaret Maclennan, Paul W. Stratford and Deborah J. Cook

Am J Crit Care 2009;18 58-64 10.4037/ajcc2009757
©2009 American Association of Critical-Care Nurses
Published online http://ajcc.aacnjournals.org/

Personal use only. For copyright permission information:
http://ajcc.aacnjournals.org/cgi/external_ref?link_type=PERMISSIONDIRECT

Subscription Information
http://ajcc.aacnjournals.org/subscriptions/

Information for authors
http://ajcc.aacnjournals.org/misc/ifora.xhtml

Submit a manuscript
http://www.editorialmanager.com/ajcc

Email alerts
http://ajcc.aacnjournals.org/subscriptions/etoc.xhtml