OPEN

Inter-Rater Reliability and Impact of Disagreements on Acute Physiology and Chronic Health Evaluation IV Mortality Predictions

Michelle Simkins, RN, MPH[1]; Ayesha Iqbal, MD, MPH[1]; Audrey Gronemeyer, MPH[1]; Lisa Konzen, RN, BSN, MA[2]; Jason White, RN, BSN[2]; Michael Koenig, RN[2]; Chris Palmer, MD[3]; Paul Kerby, MD[3]; Sara Buckman, MD, PharmD[4]; Vladimir Despotovic, MD[5]; Christine Hoehner, MSPH, PhD[1]; Walter Boyle, MD[3]

Objectives: Acute Physiology and Chronic Health Evaluation is a well-validated method to risk-adjust ICU patient outcomes. However, predictions may be affected by inter-rater reliability for manually entered elements. We evaluated inter-rater reliability for Acute Physiology and Chronic Health Evaluation IV manually entered elements among clinician abstractors and assessed the impacts of disagreements on mortality predictions.

Design: Cross-sectional.

Setting: Academic medical center.

Subjects: Patients admitted to five adult ICUs.

Interventions: None.

Measurements and Main Results: Acute Physiology and Chronic Health Evaluation IV manually entered elements were abstracted from a selection of charts ($n = 41$) by two clinician "raters" trained in Acute Physiology and Chronic Health Evaluation IV methodology. Rater agreement (%) was determined for each manually entered element, including Acute Physiology and Chronic Health Evaluation diagnosis, Glasgow Coma Scale score, admission source, chronic conditions, elective/emergency surgery, and ventilator use. Cohen's kappa (K) or intraclass correlation coefficient was calculated for nominal and continuous manually entered elements, respectively. The impacts of manually entered element choices on Acute Physiology and Chronic Health Evaluation IV mortality predictions were computed using published Acute Physiology and Chronic Health Evaluation IV equations, and observed to expected hospital mortality ratios were compared between rater groups. The majority of manually entered element inconsistency was due to disagreement in choice of Glasgow Coma Scale (63.8% agreement, 0.83 intraclass correlation coefficient), Acute Physiology and Chronic Health Evaluation diagnosis (68.3% agreement, 0.67 kappa), and admission source (90.2% agreement, 0.85 kappa). The difference in predicted mortality between raters related to Glasgow Coma Scale disagreements was significant (observed to expected mortality ratios for Rater 1 [1.009] vs Rater 2 [1.134]; $p < 0.05$). Differences related to Acute Physiology and Chronic Health Evaluation diagnosis or admission source disagreements were negligible. The new "unable to score" choice for Glasgow Coma Scale was used for 18% of Glasgow Coma Scale measurements but accounted for 63% of "major" Glasgow Coma Scale disagreements, and 50% of the overall difference in Acute Physiology and Chronic Health Evaluation-predicted mortality between raters.

Conclusions: Inconsistent use among raters of the new "unable to score" choice for Glasgow Coma Scale introduced in Acute Physiology and Chronic Health Evaluation IV was responsible for important decreases in both Glasgow Coma Scale and Acute Physiology and Chronic Health Evaluation IV mortality prediction

[1] Center for Clinical Excellence, BJC HealthCare, St. Louis, MO.

[2] Barnes Jewish Hospital, St. Louis, MO.

[3] Department of Anesthesiology, Washington University School of Medicine, St. Louis, MO.

[4] Department of Surgery, Washington University School of Medicine, St. Louis, MO.

[5] Department of Internal Medicine, Washington University School of Medicine, St. Louis, MO.

This work was performed at BJC HealthCare and Washington University School of Medicine, St. Louis, MO.

The authors have disclosed that they do not have any potential conflicts of interest.

For information regarding this article, E-mail: boylew@wustl.edu

Crit Care Expl 2019; 1:e0059

DOI: 10.1097/CCE.0000000000000059

reliability in our study. A Glasgow Coma Scale algorithm we developed after the study to improve reliability related to use of this new "unable to score" choice is presented.

Key Words: Acute Physiology and Chronic Health Evaluation; Glasgow Coma Scale; hospital mortality; intensive care units; outcome assessment (healthcare); predictive scoring systems; reproducibility of results; statistical models; telemedicine/tele-intensive care unit

Several scoring systems have been developed to measure disease severity and predict outcomes among ICU patients for quality benchmarking and outcomes research (1–5). Among these, the Acute Physiology and Chronic Health Evaluation (APACHE) series is widely used, with the most recent revision (APACHE IV) accounting for more than half of the estimated use of ICU scoring systems in the United States in 2012 (1). APACHE methodology uses age, and abnormal physiologic and laboratory values measured during the "APACHE day," as well as several manually entered elements (MEEs)—including APACHE ICU admission diagnosis, presence of certain chronic health conditions, admission source, elective/emergency surgery, ventilator use, and Glasgow Coma Scale (GCS) score—in the predictive model. APACHE then provides validated risk-adjusted predictions of outcomes for ICU patients including mortality and length of stay (4, 5).

Since first published in 1982, APACHE has undergone three major revisions (4–8). The latest version, APACHE IV, uses 142 variables in the predictive model and provides validated outcome predictions for 116 disease categories, compared with 78 for APACHE III (4, 5, 8). A number of studies have demonstrated that APACHE IV has superior accuracy for predicting outcomes of ICU patients when compared with other scoring systems (9–14). Consistency of APACHE data abstraction and the impact of rater disagreements on reliability of predictions have also been studied, but only with earlier versions of APACHE (15–21). MEE reliability has not been investigated using APACHE IV, nor has the potential impact of APACHE IV MEE disagreements on the reliability APACHE IV predictions been systematically evaluated.

We began using APACHE IV with implementation of an ICU telemedicine program which uses APACHE IV for outcomes benchmarking and comparisons between programs. Given the importance of MEE consistency to prediction reliability, we evaluated inter-rater reliability for MEEs among clinician chart abstractors and determined the impact of disagreements on APACHE IV group-level mortality predictions using published APACHE IV equations.

MATERIALS AND METHODS

Setting/Study Population

Prior to implementation of the ICU telemedicine program in January 2016, training in APACHE methodology for clinician chart abstractors was conducted over 3 months, between October 2015 and December 2015. The training included completion of APACHE IV training modules and case studies and familiarization with procedures developed to standardize APACHE IV data entry. This reliability assessment was then conducted 3 months following implementation, between March 2016 and July 2016. The goal was to evaluate consistency of APACHE IV MEE data abstraction and to determine impacts of MEE disagreements on mortality predictions. Nine clinician data abstractors (eight registered nurses and one physician) participated in the evaluation. Chart abstraction data were collected for patients admitted to the five adult ICUs in the program, with a quasi-random sampling of charts to ensure equal distribution between ICUs, and daytime versus nighttime admissions.

Data Collection

This study was reviewed by the Washington University Institutional Review Board (IRB) and IRB approval was waived, with no requirement for informed consent. APACHE IV MEE data were collected in a web-based application for online databases (Research Electronic Data Capture, REDCap.org). MEE data included APACHE admission diagnosis (453 choices in 116 disease categories); total GCS (range 3–15) and scores for each of the three GCS components (eye, verbal, motor), or "unable to score (GCS) due to medications"; admission source (eight categories); emergency surgery (yes/no); any of nine chronic conditions (yes/no); and mechanical ventilation (yes/no).

Forty-one charts were selected for inclusion in this study, which satisfied the minimum sample size recommended for detecting a statistically significant difference for a dichotomous variable (22). All charts were abstracted by two clinician "raters" whose data were blinded from each other. "Rater 1" was the admitting clinician who abstracted APACHE MEEs from information in the electronic medical record (EMR) in real time during the APACHE day, or the following day for patients admitted after 6 PM. "Rater 2" abstracted information from the EMR retrospectively. To ensure that all the raters used information from the same time period, both rater groups were instructed to only consider information in the EMR available during the APACHE day (extending to midnight on the day of admission, or to midnight the following day for ICU admissions after 4 PM).

Measures/Analysis

All analyses were conducted using SAS Software 9.4 (SAS Institute, Cary, NC). Percent agreement among raters was calculated for each MEE where agreement was defined as rater 1 and rater 2 selecting the same option with a few specific exceptions: 1) Admission sources of operating or recovery room were considered in agreement; 2) APACHE diagnoses were considered in agreement as long as both diagnoses were from the same diagnosis category ($n = 116$) (i.e., both share the same APACHE diagnosis coefficient); and 3) Use of "unable to score due to medications" by only one rater was considered a "major" disagreement in total GCS irrespective of the total GCS (sum of GCS components) recorded by the other rater. (Note: The "unable to score" choice assigns a normal GCS [15] for the APACHE IV Acute Physiology Score [APS] and APACHE IV composite GCS coefficient, with no GCS contribution to predicted mortality; but when "unable to score" is recorded, a separate coefficient is inserted in the APACHE IV equations that result in a higher predicted mortality [4].—We also completed simulations using published APACHE IV equations [www.https://

intensivecarenetwork.com/Calculators/Files/Apache4.html] to determine which GCS score was equivalent to “unable to score” based on morality predictions. Specifically, we entered data for several ICU patients in the APACHE IV calculator holding all elements constant except GCS. A comparison of mortality predictions between selecting “unable to score” versus each possible GCS numeric value was then completed—see *Results* and *Discussion* sections). For the GCS components, “unable to score” by one rater was considered a disagreement if the other rater scored that component below the maximum.

MEE agreement (%) was described using the adjectival ratings of Landis and Koch (23): 80% to 100% (“almost perfect to perfect”); 60% to 80% (“substantial”); 40% to 60% (“moderate”); 20% to 40% (“fair”); and 0% to 20% (“poor”). To account for chance agreement, Cohen’s kappa or intraclass correlation coefficient (ICC) was calculated for nominal or continuous MEEs, respectively, with 1.0 representing perfect agreement (24).

The impacts of MEE disagreements between raters on the observed to APACHE IV-predicted (expected) hospital mortality ratio (O:E mortality ratio) were computed using published APACHE IV equations (www.https://intensivecarenetwork.com/Calculators/Files/Apache4.html). Effects of MEE disagreements on O:E mortality ratios were each evaluated independently—by holding the other MEEs constant (at the value recorded by rater 1). Composite effects of MEE disagreements on O:E mortality ratios were also evaluated using the composite MEE values recorded by each rater. Statistical analyses of differences in O:E mortality ratios between rater groups were accomplished using bootstrap resampling methodology.

RESULTS

As shown in **Figure 1**, agreement between Rater 1 and Rater 2 was “substantial” for APACHE diagnosis (68.3% agreement, 0.67 kappa) and GCS (63.4%, ICC 0.83), while agreement was “almost perfect to perfect” for admission source (90.2% agreement, 0.85 kappa), chronic conditions composite (96.5% agreement, 0.68 kappa), elective surgery (100% agreement, 1.0 kappa), and ventilator use (100% agreement, 1.0 kappa)

Of the 13 APACHE diagnosis category disagreements, eight (62%) represented disagreements within the same organ system, while the remaining five (38%) represented choices from different organ systems, or disagreements in surgical versus medical diagnoses, or both.

Of the 15 disagreements in total GCS, seven (46%) represented “minor” (one point) GCS disagreements. Of the remaining eight “major” GCS disagreements, five involved disagreements in use of the new “unable to score due to medications” choice for GCS introduced with APACHE IV (4, 5). This “unable to score” choice was used for 15 of 81 (18%) GCS measurements recorded by raters, including nine of 41 values recorded by Rater 1 (22%), and six of 41 values recorded by Rater 2 (15%). As shown in Figure 1, agreement for the GCS components was better than that for total GCS, with the motor component highest (82.9%), and verbal lowest (70.7%). “Unable to score” disagreements accounted for five of seven disagreements for the motor component (71.4%), five of eight disagreements for the eye component (62.5%), and five of 12 disagreements for the verbal component (41.7%).

The impacts of MEE disagreements in APACHE diagnosis, GCS, and admission source, which contributed the majority of MEE inconsistency, as well as the composite effect of MEE disagreements, on O:E mortality ratios as determined using the published APACHE IV equations are shown in **Figure 2**. Despite 31.7% disagreement between raters in choice of APACHE diagnosis category, the diagnosis disagreements had a negligible effect on group-level APACHE mortality predictions (O:E hospital mortality for Rater 1 [1.009] vs Rater 2 [1.007]). In contrast, the 36.4% disagreement in GCS between raters resulted in a large and significant difference in group-level APACHE mortality predictions (O:E for Rater 1 [1.009] vs Rater 2 [1.134]; $p < 0.05$). Disagreement in use of the “unable to score” choice, which accounted for 63% of the “major” GCS disagreements, accounted for 50% of the overall difference in predicted mortality between rater groups related to GCS. Notably, one “unable to score” disagreement produced only a small difference in predicted mortality between raters (0.25%) when the other rater recorded a numeric GCS of 10. The analysis of the impact of the “unable to score” choice on predicted mortality done using simulations with the published APACHE IV

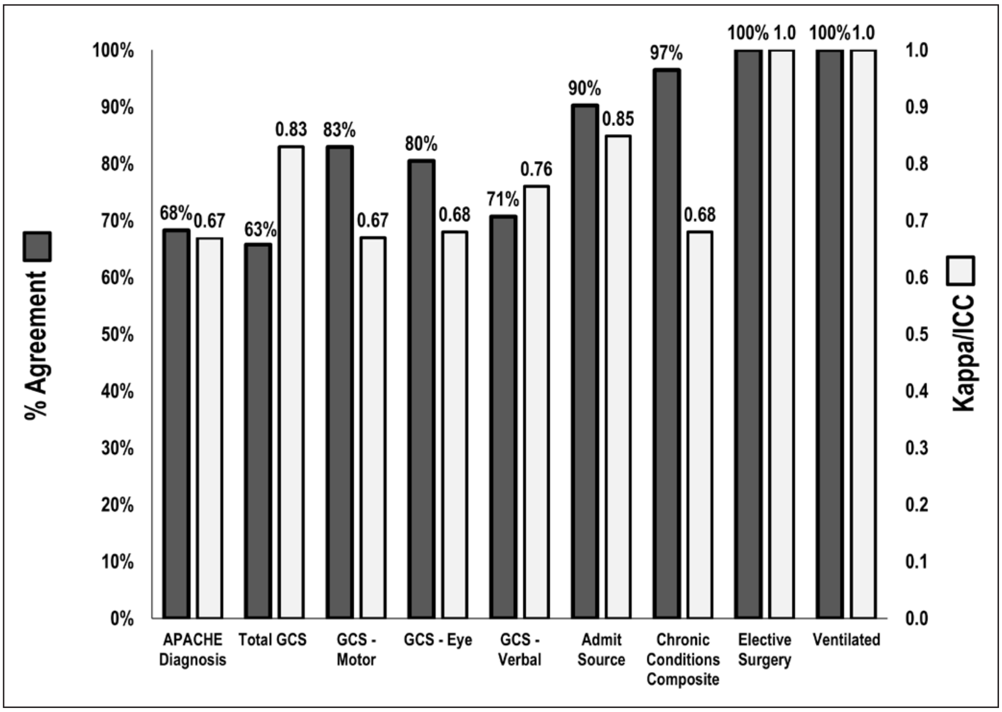

Figure 1. Inter-rater reliability for Acute Physiology and Chronic Health Evaluation (APACHE) manually entered elements. % agreement is shown together with the Kappa statistic for nominal variables and interclass correlation coefficient (ICC) for continuous variables (see Text). GCS = Glasgow Coma Scale.

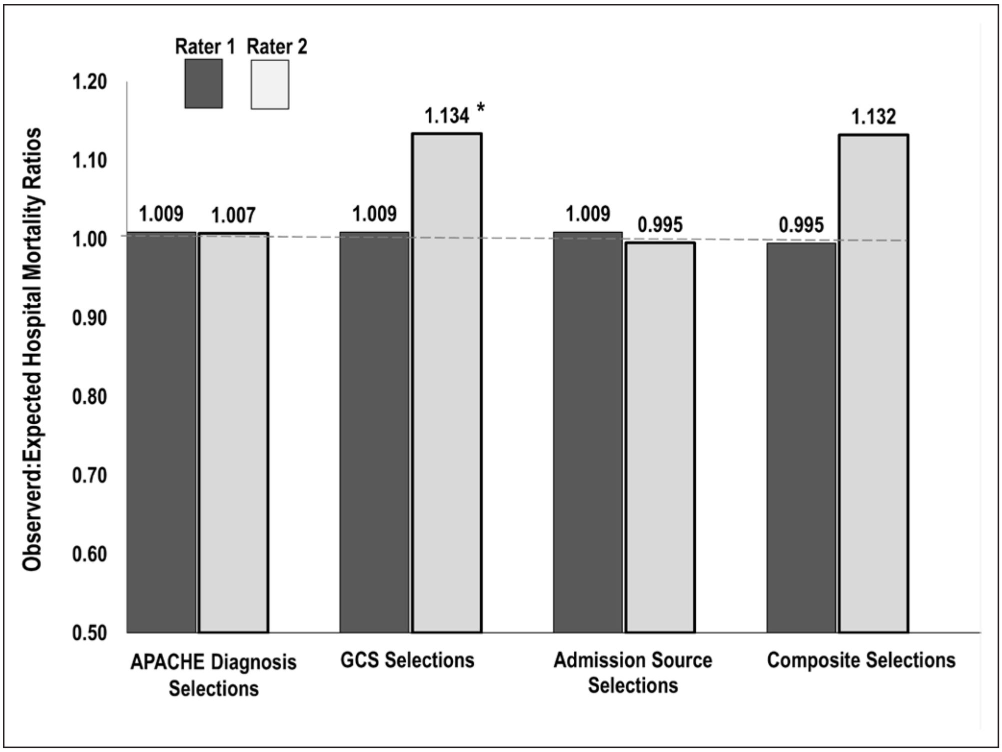

Figure 2. Impact of manually entered element (MEE) disagreements between raters on Acute Physiology and Chronic Health Evaluation (APACHE) IV group-level mortality predictions. Calculated APACHE IV observed to expected mortality ratios for the two rater groups are shown for each of the three MEEs evaluated independently, as well as the composite effects of MEE disagreements on the mortality predictions. *Significant difference between rater 1 versus rater 2 ($p < 0.05$). GCS = Glasgow Coma Scale.

calculator (www.https://intensivecarenetwork.com/Calculators/Files/Apache4.html) demonstrated that: 1) the APACHE IV "unable to score" choice coefficient had an effect on predicted mortality that was closely approximated by a numeric GCS of 9–10 (with 13–15 APS points); and 2) compared to the earlier practice of recording a normal GCS (15) when GCS assessment was not possible (4, 5), the "unable to score" choice resulted in an increase in individual predicted mortality by as much as 20%.

As shown in Figure 2, the 9.8% disagreement in admission source had only a small effect on group-level APACHE predicted mortality (O:E for Rater 1 [1.009] vs Rater 2 [0.995]). The composite effect of MEE disagreements on group-level APACHE predicted mortality between raters appeared to mirror the effect of the GCS disagreements alone (O:E for Rater 1 [0.995] vs Rater 2 [1.132]), although this composite difference did not reach statistical significance (Fig. 2).

DISCUSSION

Agreement for MEEs in this APACHE IV reliability study was "substantial" to "perfect" between raters, consistent with prior studies using earlier versions of APACHE (15–21). APACHE diagnosis had the lowest reliability based on the kappa statistic, which was not unexpected given the large number of diagnosis category choices in APACHE IV (*n* = 116), and the subjectivity in choosing a single APACHE diagnosis in patients with multiple problems. There were a few larger differences (> 10%) in individual predicted mortality related to rater diagnosis choice disagreement, but the overall differences were well-balanced and did not impact group-level APACHE mortality predictions, consistent with earlier reports (15–21). Similar to prior studies, we found little disagreement in less subjective MEEs (15, 16, 18).

Inconsistency between raters in GCS was also not unexpected given the know subjectivity of GCS (25), and the level of GCS agreement we observed was similar to that reported using prior versions of APACHE (15, 17, 20, 21). In contrast to these earlier studies, however, we found that GCS disagreements had a significant impact on APACHE IV group-level mortality predictions. GCS is a required element of APACHE, and despite its known subjectivity, has the largest potential impact on APACHE mortality predictions. A GCS of 3 contributes 48 points to the APACHE IV APS, representing 19% of the 252 point maximum (4, 5), similar to 17% in APACHE II–III (7, 8, 26). New in APACHE IV, and highlighted by our study, is the "unable to score due to medications" choice for GCS that was introduced to "reduce predictive inaccuracies caused by defaulting GCS to normal (15) when assessment was not possible." (4, 5) This new choice was used frequently (18% of measurements) and was a major source of GCS inconsistency that significantly affected mortality predictions in our study. When considered in the context of prior studies with earlier versions of APACHE that did not have an "unable to score" choice, and did not demonstrate any difference in APACHE predictions related to MEE disagreements (15, 17), our finding suggests the new "unable to score" choice in APACHE IV has amplified the impact of GCS inconsistency on the reliability of APACHE mortality predictions.

Differences in our analysis may also have contributed to different conclusions regarding the impact of GCS inconsistency in comparison to earlier studies (15–21). In particular, we only found a significant effect when GCS disagreements were considered independent of other MEE disagreements, and consistent with earlier reports, we did not find a significant effect of composite MEE disagreements on predicted mortality (15, 17). However, we cannot agree with the earlier conclusion that MEE variability is sufficiently random and offsetting as to have no significant impact on group-level predictions (15). The effect of composite MEE disagreements on predicted mortality we observed, while not significant in this relatively small study, was nearly identical to, and appeared to reflect, the effect of the GCS disagreements (Fig. 2). At a minimum, our findings demonstrate the importance of GCS reliability and again highlight the potential impact of inconsistency in use of the new "unable to score" choice in APACHE IV on both GCS and mortality prediction reliability.

Lack of understanding regarding the appropriate use of the new “unable to score” choice, as well as misconceptions regarding its impact, appear to be likely contributors to its inconsistent use in our study. A commonly stated misconception among data abstractors was that “unable to score” was approximately equivalent to the prior practice of selecting a GCS of 15 in sedated or anesthetized patients, which is true for APACHE APS calculations, but not for mortality predictions (4, 5). Accordingly, we developed a GCS algorithm after the study that incorporates a standardized approach to use of the new “unable to score” choice as we believe it was intended (**Fig. 3**). Per the algorithm, “unable to score” is used in patients receiving CNS depressing or neuromuscular blocking medications when there is either no recent (within 12 hr) assessment to allow GCS to be scored prior to receiving such medication(s), or a reasonable likelihood that GCS has been affected by the condition or the procedure since the prior assessment. An exception is made for postoperative patients where a normal GSC (15) is assumed in the absence of a recent prior assessment, again unless there is a reasonable likelihood that GCS has been affected by the patient’s condition or procedure, in which case the “unable to score” choice is used. The algorithm also incorporates use of the “modified” verbal score for awake patients who are unable to provide a verbal response (e.g., intubated patients). Further reliability studies will be needed to demonstrate that this new tool positively impacts GCS and APACHE prediction reliability. However, since instituting this algorithm in our program, our clinician abstractors report improved clarity in making GCS determinations, and in use of the “unable to score” choice. We are thus providing the algorithm for potential use by others using APACHE IV methodology, or other outcome measurements that use GCS.

As indicated in Results, our patient data and simulations using the APACHE IV calculator indicate that the impact of “unable to score” on mortality predictions approximates a numeric GCS of 9–10 (depending on the scoring of the individual components). Interestingly, our group recently contacted Philips Healthcare regarding their mortality prediction model (embedded in their ICU telemedicine software), which required a numeric GCS.

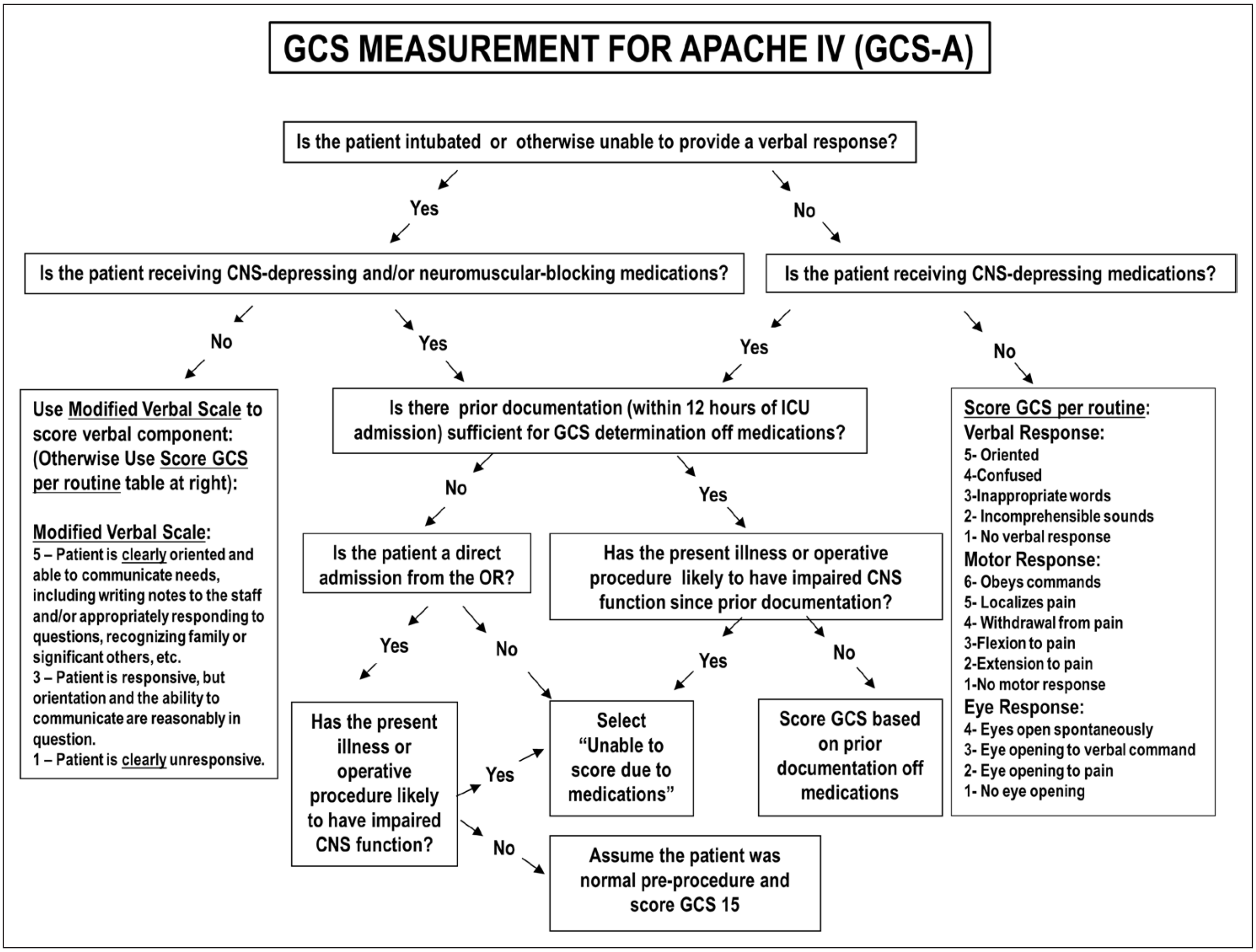
GCS MEASUREMENT FOR APACHE IV (GCS-A)
Is the patient intubated or otherwise unable to provide a verbal response?
Yes
No
Is the patient receiving CNS-depressing and/or neuromuscular-blocking medications?
Is the patient receiving CNS-depressing medications?
No
Yes
Yes
No
Use Modified Verbal Scale to score verbal component: (Otherwise Use Score GCS per routine table at right):
Modified Verbal Scale:
5 – Patient is clearly oriented and able to communicate needs, including writing notes to the staff and/or appropriately responding to questions, recognizing family or significant others, etc.
3 – Patient is responsive, but orientation and the ability to communicate are reasonably in question.
1 – Patient is clearly unresponsive.
Is there prior documentation (within 12 hours of ICU admission) sufficient for GCS determination off medications?
No
Yes
Is the patient a direct admission from the OR?
Has the present illness or operative procedure likely to have impaired CNS function since prior documentation?
Yes
No
Yes
No
Has the present illness or operative procedure likely to have impaired CNS function?
Yes
No
Select “Unable to score due to medications”
Score GCS based on prior documentation off medications
Assume the patient was normal pre-procedure and score GCS 15
Score GCS per routine:
Verbal Response:
5- Oriented
4-Confused
3-Inappropriate words
2- Incomprehensible sounds
1- No verbal response
Motor Response:
6- Obeys commands
5- Localizes pain
4- Withdrawal from pain
3-Flexion to pain
2-Extension to pain
1-No motor response
Eye Response:
4- Eyes open spontaneously
3- Eye opening to verbal command
2- Eye opening to pain
1- No eye opening

Figure 3. Glasgow Coma Scale (GCS) algorithm developed for Acute Physiology and Chronic Health Evaluation (APACHE) IV incorporating a standardized approach for recording GCS and for use of the new “unable to score due to medications” choice. OR = operating room.

Following their independent investigation, the Philips group concluded "…the mortality risk associated with the 'unable to score due to meds' was very similar to that of a GCS of 8," and "unable to score" is now converted to GSC 8 in their model (O. Badawi, personal communication, 2018). The Philips data provide additional validation of the impact assigned to the "unable to score" choice on predicted mortality, and further suggest both the APACHE IV and Philips mortality prediction models are similarly calibrated with respect to GCS. These findings also have potential implications for calibration of "unable to score" in other prediction models that use GCS.

It is important to note a few limitations of our study which could impact broader applicability of our findings. First, although our study had an acceptable sample size for reliability testing, the number of charts abstracted was relatively small, and from one institution. Additionally, our population included a high percentage of intubated and sedated patients in whom "unable to score" may be used more frequently than in other ICU patient populations. It is also important to also note that Rater 1 did chart abstraction in real time, while chart abstraction by the Rater 2 group was done retrospectively. Although we attempted to control for this by confining the epoch for review to the APACHE day for both rater groups, the timing of data abstraction could have contributed to the observed GCS inconsistency. Given the importance of GCS reliability, perhaps APACHE IV GCS data abstraction timing should be studied further. It is important to note, however, that use of the "unable to score" choice, which we identified as a major source of GCS disagreements that affected mortality predictions, appeared well-balanced between the two rater groups in our study.

CONCLUSIONS

Our study addresses an important gap in research related to APACHE IV reliability. To our knowledge, this is the first study to systematically evaluate APACHE IV MEE disagreements and their potential impact on APACHE IV mortality predictions. Our study demonstrates the previously well-described subjectivity in GCS determinations, but goes an important step further to highlight the potential importance of the new "unable to score" choice for GCS on APACHE IV mortality prediction reliability. We provide a GCS algorithm to improve GCS reliability, particularly related to use of the new "unable to score" choice available when using APACHE IV methodology.

ACKNOWLEDGMENTS

We thank the ICU telemedicine program staff for their significant contributions to this work. We also thank Audrey Herring and the clinical epidemiology team in the BJC Center for Clinical Excellence for their helpful edits of this article.

REFERENCES

1. Breslow MJ, Badawi O: Severity scoring in the critically ill: Part 1–interpretation and accuracy of outcome prediction scoring systems. *Chest* 2012; 141:245–252
2. Moreno RP, Metnitz PG, Almeida E, et al; SAPS 3 Investigators: SAPS 3–from evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission. *Intensive Care Med* 2005; 31:1345–1355
3. Salluh JI, Soares M: ICU severity of illness scores: APACHE, SAPS and MPM. *Curr Opin Crit Care* 2014; 20:557–565
4. Zimmerman JE, Kramer AA, McNair DS, et al: Acute Physiology and Chronic Health Evaluation (APACHE) IV: Hospital mortality assessment for today's critically ill patients. *Crit Care Med* 2006; 34:1297–1310
5. Zimmerman JE, Kramer AA, McNair DS, et al: Intensive care unit length of stay: Benchmarking based on Acute Physiology and Chronic Health Evaluation (APACHE) IV. *Crit Care Med* 2006; 34:2517–2529
6. Knaus WA, Draper EA, Wagner DP, et al: Evaluating outcome from intensive care: A preliminary multihospital comparison. *Crit Care Med* 1982; 10:491–496
7. Knaus WA, Draper EA, Wagner DP, et al: APACHE II: A severity of disease classification system. *Crit Care Med* 1985; 13:818–829
8. Knaus WA, Wagner DP, Draper EA, et al: The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. *Chest* 1991; 100:1619–1636
9. Kramer AA, Higgins TL, Zimmerman JE, et al: Comparison of the Mortality Probability Admission Model III, National Quality Forum, and Acute Physiology and Chronic Health Evaluation IV hospital mortality models: Implications for national benchmarking*. *Crit Care Med* 2014; 42:544–553
10. Sánchez-Casado M, Hostigüela-Martín VA, Raigal-Caño A, et al: Predictive scoring systems in multiorgan failure: A cohort study. *Med Intensiva* 2016; 40:145–153
11. Juneja D, Singh O, Nasa P, et al: Comparison of newer scoring systems with the conventional scoring systems in general intensive care population. *Minerva Anestesiol* 2012; 78:194–200
12. Keegan MT, Gajic O, Afessa B: Comparison of APACHE III, APACHE IV, SAPS 3, and MPM0III and influence of resuscitation status on model performance. *Chest* 2012; 142:851–858
13. Kuzniewicz MW, Vasilevskis EE, Lane R, et al: Variation in ICU risk-adjusted mortality: Impact of methods of assessment and potential confounders. *Chest* 2008; 133:1319–1327
14. Vasilevskis EE, Kuzniewicz MW, Cason BA, et al: Mortality probability model III and simplified acute physiology score II: Assessing their value in predicting length of stay and comparison to APACHE IV. *Chest* 2009; 136:89–101
15. Chen LM, Martin CM, Morrison TL, et al: Interobserver variability in data collection of the APACHE II score in teaching and community hospitals. *Crit Care Med* 1999; 27:1999–2004
16. Damiano AM, Bergner M, Draper EA, et al: Reliability of a measure of severity of illness: Acute physiology of chronic health evaluation—II. *J Clin Epidemiol* 1992; 45:93–101
17. Kho ME, McDonald E, Stratford PW, et al: Interrater reliability of APACHE II scores for medical-surgical intensive care patients: A prospective blinded study. *Am J Crit Care* 2007; 16:378–83
18. Polderman KH, Girbes AR, Thijs LG, et al: Accuracy and reliability of APACHE II scoring in two intensive care units: Problems and pitfalls in the use of APACHE II and suggestions for improvement. *Anaesthesia* 2001; 56:47–50
19. Polderman KH, Jorna EM, Girbes AR: Inter-observer variability in APACHE II scoring: Effect of strict guidelines and training. *Intensive Care Med* 2001; 27:1365–1369
20. Wenner JB, Norena M, Khan N, et al: Reliability of intensive care unit admitting and comorbid diagnoses, race, elements of Acute Physiology and Chronic Health Evaluation II score, and predicted probability of mortality in an electronic intensive care unit database. *J Crit Care* 2009; 24:401–407
21. Holt AW, Bury LK, Bersten AD, et al: Prospective evaluation of residents and nurses as severity score data collectors. *Crit Care Med* 1992; 20:1688–1691

22. Sim J, Wright CC: The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Phys Ther* 2005; 85:257–268

23. Landis JR, Koch GG: The measurement of observer agreement for categorical data. *Biometrics* 1977; 33:159–174

24. Szklo M, Nieto, FJ: Quality assurance and control. *In*: Epidemiology: Beyond the Basics. Third Edition. Johnson M, Reilly T (Eds). Burlington, VT, Jones & Barlett Learning, 2014, pp 313–366

25. Reith FC, Van den Brande R, Synnot A, et al: The reliability of the Glasgow Coma Scale: A systematic review. *Intensive Care Med* 2016; 42:3–15

26. Livingston BM, Mackenzie SJ, MacKirdy FN, et al: Should the pre-sedation Glasgow Coma Scale value be used when calculating Acute Physiology and Chronic Health Evaluation scores for sedated patients? Scottish Intensive Care Society Audit Group. *Crit Care Med* 2000; 28:389–394