# Q3: An ICU Acuity Score for Exploring the Effects of Changing Acuity Throughout the Stay

Stephen E. Brossette MD, PhD[1], Ning Zheng PhD[1], Daisy Y. Wong PhD[1], Patrick A. Hymel Jr. MD[1]

[1] Indicator Sciences, LLC

Corresponding author:
Stephen E. Brossette, MD, PhD
E-mail: sbrossette@indisci.com

keywords: critical care, acuity scores, outcomes, electronic health record

# Abstract

## Intro

We develop a straightforward ICU acuity score (Q3) that is calculated every 3 hours throughout the first 10 days of the ICU stay. Q3 uses components of the Oxford Acute Severity of Illness Score (OASIS) and incorporates a new component score for vasopressor use. In well-behaved models of ICU mortality, the marginal effects of Q3 are significant across the first 10 days of the ICU stay. In separate models, Q3 has significant effects on ICU remaining length of stay. The score has implications for work that seeks to explain modifiable mechanisms of changing acuity during the ICU stay.

## Methods

From the MIMIC-III database, select ICU stays from 5 adult ICUs were partitioned into consecutive 3-hour segments. For each segment, the number of vasopressors administered and all 10 OASIS component scores were computed. Models of ICU mortality were estimated. OASIS component effects were examined, and vasopressor count bins were weighted. Q3 was defined as the sum of 8 retained OASIS components and a new weighted vasopressor score. Models of ICU mortality quadratic in Q3 were estimated for each of the first 10 ICU days and were subjected to segment-level, location-specific tests of discrimination and calibration on newer ICU stays. Marginal effects of Q3 were computed at different levels of Q3 by ICU day, and average marginal effects of Q3 were computed at each location by ICU day. ICU remaining length of stay (LOS) models were also estimated and the effects of Q3 were similarly examined.

## Results

Daily ICU mortality models using Q3 show no evidence of misspecification (Pearson-Windmeijer $p>0.05$, Stukel $p>0.05$), discriminate well in all ICUs over the first 10 days (AUROC ~ 0.72 – 0.85), and are generally well calibrated (Hosemer-Lemeshow $p>0.05$, Spiegelhalter's z $p>0.05$). A one-unit increase in Q3 from typical levels (Q3=15) affects the odds of ICU mortality by a factor of 1.14 to 1.20, depending on ICU day ($p<0.001$), and the ICU remaining LOS by 5.8 to 9.6% ($p<0.001$). On average, a one-unit increase in Q3 increases the probability of ICU mortality by 1 to 2 percentage points depending on location and ICU day, and ICU remaining LOS by 5 to 10 hours depending on location and ICU day.

## Conclusion

Q3 significantly affects ICU mortality and ICU remaining LOS in different ICUs across the first 10 days of the ICU stay. Depending on location and ICU day, a one-unit increase in Q3 increases the probability of ICU mortality by or 1-2 percentage points and ICU remaining LOS by 5 to 10 hours. Unlike static acuity scores or those updated infrequently, Q3 could be used in explanatory models to help elucidate mechanisms of changing ICU acuity.

# Introduction

## Background and Significance

Many ICU acuity scores have been developed [1-7]. They equate acuity with odds of mortality (in-ICU or in-hospital) and are used in models that discriminate well and are sometimes well calibrated. Acuity scores differ in variables used, but most use vital signs, laboratory results, medications, ventilator status, comorbidities, and patient demographics. Established ICU acuity scores include APACHE, SAPS, MPM, SOFA, and OASIS, amongst others. Performance is typically evaluated at the end of the first ICU day (except SOFA which is also evaluated every subsequent 48 hours) with discrimination evaluated by AUROC/c-statistics, and calibration examined via Hosemer-Lemeshow (HL) tests.

The Oxford Acute Severity of Illness Score (OASIS) is a relatively new ICU acuity score. It uses less data than others (10 variables), does not depend on laboratory results which can lag in acquisition and reporting, and performs about as well as other acuity scores [7]. It was designed to be more computable and timelier than other scores, but like most other scores, has only been evaluated at the end of the first ICU day.

Real-time models of ICU acuity have also been developed [8-10]. They typically use many clinical variables instead of a single summarized acuity score, and rely on machine learning techniques to capture complex, non-linear relationships between explanatory variables and the outcome. While they typically perform well, machine learning models are (so far) uninterpretable, and the marginal effects of key variables on the odds of mortality are complex and unknown. It is also notable that in a

study by Johnson and Mark [10], machine learning techniques only slightly outperformed logistic regression models, which are interpretable.

In this study we construct a new ICU acuity score (Q3) that uses a subset of OASIS component scores and incorporates a new component score based on vasopressor use. Q3 is computed every 3 hours throughout the ICU stay. We show that straightforward models quadratic in Q3 are well-behaved over the first 10 ICU days across different ICUs, and that the marginal effects of Q3 on ICU mortality and ICU remaining LOS are significant and relatively stable.

## Materials and Methods

### Data

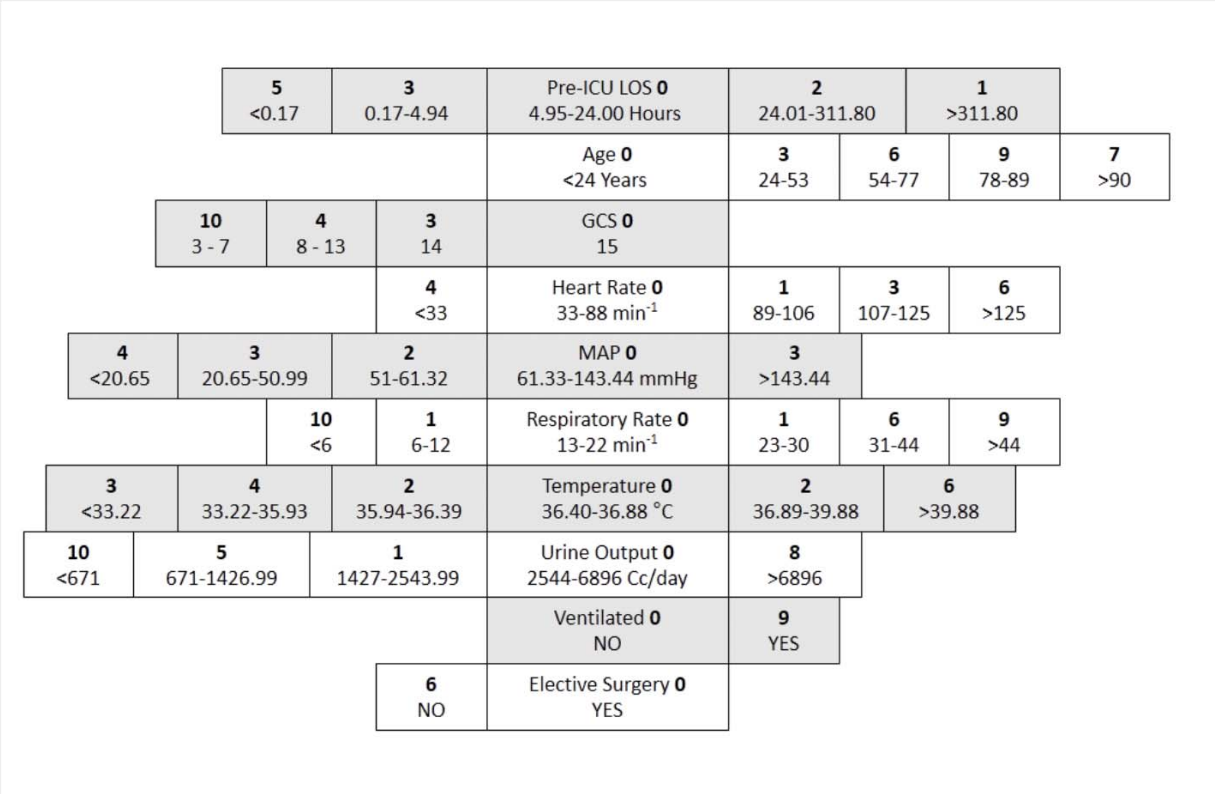
We analyzed data from the MIMIC-III database, version 1.4. MIMIC-III is comprised of de-identified data from over forty thousand patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012 [11-12]. It contains data from Carevue and Metavision ICU systems, the laboratory information system, and the hospital EHR. Carevue data are from years 2001-2008 and Metavision data are from years 2008-2012, with very little overlap.

We used adult (>=18yo) ICU stays from the MICU (medical), SICU (surgical), CCU (coronary care), CSRU (cardiac surgery recovery), and TSICU (trauma surgical) (n=52,061). Each stay was divided into contiguous, non-overlapping 3-hour *segments*, indexed from 0.

**Computing Q3 component scores**

For each 3-hour segment, we computed the number of distinct vasopressors administered along with all 10 OASIS component scores. The Oxford Acute Severity of Illness Score (OASIS) is the sum of the first-day maximum scores of 10 component scores, described by Johnson, et. al [7]. Using the same component bins and weights (Fig 1), we computed each component score at the end of each segment for 80 segments (10 days) using the latest-available data, without considering relative score values over time. Specifically, each component score used the latest data over the previous 24 hours, except for *preiculos*, *age*, and *electivesurg*, which were constant throughout the stay. If no component data existed within 24 hours, the component score was set to a default value of 0. For all components except *urineout*, the implementation of the latest-data logic was straightforward. For *urineout*, the 24-hour rate of urine output, the following logic was used. At each urine output event, a 24-hour rate of urine output was computed by dividing the volume of urine collected since the last event by the number of elapsed hours since that event or the start of the ICU stay, if no event existed. This rate was assigned to the time window from the current output event back to the previous event (or start of the ICU stay). Then for each segment, *urineout* was computed as the average of all 24-hour rates assigned to time windows that overlapped with the segment. If the end of a segment occurred after the last urine output event, *urineout* was set the prior value of *urineout*.

Figure 1: Component bins and weights for the Oxford Acute Severity of Illness Score (OASIS) [7]




Vasopressors were comprised of norepinephrine, epinephrine, vasopressin, dobutamine, dopamine, milrinone, and phenylephrine. For each segment, the number of distinct vasopressors used was binned into none, one, two, and 3 or more (Table 1).

Table 1: Vasopressor use in first 10 ICU days

| n vasopressors | % segments |
|---|---|
| 0 | 82.2 |
| 1 | 13.6 |
| 2 | 3.2 |
| >=3 | 1.1 |

**Selecting Q3 components and weighting vasopressor use bins**

The effects of OASIS components and vasopressor use on ICU mortality were examined through estimates of Equation A.

$$A)\ logit(mort) = A_0 + \boldsymbol{A_1 OasisCompScores} + \boldsymbol{A_2 vpresBins} + A_3 vpres + u$$

In this equation, *logit* is the logistic function; *mort* is an indicator for ICU mortality; ***OasisCompScores*** is a vector of the 10 OASIS component scores computed at each segment; ***vpresBins*** is a vector of 4 indicators for the number of distinct vasopressors administered in the segment (n=0,1,2,>=3); and *vpres* is the weight assigned to the vasopressor bin for the number of vasopressors administered in the segment. The four vasopressor bins were weighted by first estimating Equation A without *vpres* and using the parameter estimates of ***vpresBins*** to calculate a starting weight for each bin. Then a simple integer-grid-search around the starting weights was executed until estimates of Equation A (with *vpres*) produced *vpres* effects comparable to those of other significant components and ***vpresBins*** effects that were insignificant. Throughout, Equation A was estimated once for each of the first 10 ICU days, using only the fourth segment of each day to avoid repeated sampling of the same patients. Parameter estimates across all 10 days were examined and a subset of components with consistent and significant effects was retained. Finally, Q3 was defined as the sum of the retained components.

**Estimating the effects of Q3 on ICU mortality**

The effects of Q3 on ICU mortality were examined through estimates of Equation B.

$$B)\ logit(mort) = B_0 + B_1 Q3 + B_2 Q3^2 + \boldsymbol{B_3 Loc} + u$$

The new variable ***Loc*** is a vector of 5 location indicators for MICU, SICU, CCU, CSRU, and TSICU. Equation *B* was estimated 10 times using Carevue stays (2001-2008), once for each of the first 10 ICU days, using data from the fourth segment of each day. Pearson-Windmeijer [13] and Stukel [14] goodness-of-fit tests were used to check for sources of misspecification.

**Discrimination and calibration testing**

Discrimination is a measure of a model's ability to assign higher scores to patients with events (e.g. death) than without, and calibration is a measure of a model's ability to accurately estimate the number of events in different quantiles of probability. The 10 estimated daily models of Equation *B* (using Carevue stays, 2001-2008) were tested for discrimination and calibration on strictly newer Metavision stays (2008-2012). By ICU, the day 0 model was tested at each segment 0 through 7, the day 1 model at segments 8 through 15, etc. Discrimination was evaluated by AUROC/c-statistics and calibration by Hosemer-Lemeshow (HL) tests using 10 groups. Since HL tests become too powerful in large samples [15], some have recommended that HL tests use smaller random samples where appropriate [15-18]. Although exact sample size recommendations are unavailable, and the determination of calibration is ultimately subjective, one strategy suggested by Paul et. al. is to perform HL tests on random samples of size n=1000 and 10 groups [16], and others have employed random

sampling for HL calibration tests [17-18]. We follow this strategy and use a random sample of n=1000 observations when more than 1000 observation were available. An insignificant HL test ($p>0.05$) is suggestive of sufficient segment-level calibration. Overall model calibration was deemed sufficient if, in a single ICU, segment-level calibration results were mostly insignificant over extended periods of time. In addition to c-statistics and HL tests, Brier scores were computed for each segment and location. Brier scores simultaneously capture features of discrimination and calibration [19]. The Spiegelhalter's z-test was used to test for Brier score significance, and p-values were computed.

**Marginal and average marginal effects**

Since Equation B is quadratic in Q3, the marginal effects of Q3 depend on the level of Q3 itself as shown by the derivative of Equation B with respect to Q3: $d(\text{logit}(mort))/dQ3 = B_1 + 2B_2Q3$. Marginal effects of the 10 daily models were calculated at Q3=5, 15, 25, and 35. The average marginal effects of Q3 (AME, in probability points) were calculated daily.

**Estimating the effects of Q3 on ICU remaining LOS**

The effects of Q3 on ICU remaining LOS (rLOS) were examined through Poisson regression where the distributional parameter lambda is given by the following equation.

$$C)\ lambda = E(rLOS) = \exp\left(C_0 + C_1 Q3 + C_2 Q3^2 + \boldsymbol{C_3 Loc} + \boldsymbol{C_4 daypt} + \boldsymbol{C_5} wkend\right)$$

Here, *rLOS* is the remaining LOS (in hours) from the end of the current segment, ***daypart*** is a vector of indicators for 6 consecutive 4-hour intervals (starting from midnight) in which the segment starts, and

*wkend* is an indicator for the segment starting on Saturday or Sunday. Equation *C* was estimated once daily for each of the first 10 ICU days. Robust standard errors were employed to adjust for any overdispersion and are valid under *any* conditional variance assumption [20]. The natural log of equation C is quadratic in Q3 and its marginal effects depend on the level of Q3. They were estimated at Q3=5, 15, 25, and 35 and can be easily interpreted as proportional changes in rLOS. Average marginal effects (in hours) for each ICU day and location were also computed.

All analysis was done using Stata v16 (StataCorp. 2019. *Stata Statistical Software: Release 16*. College Station, TX: StataCorp LLC). Investigators were certified to use the MIMIC-III database. All data were previously de-identified, and the public use of MIMIC has been approved by the IRB of its hosting organization. No additional IRB approval was required. Project code is available from the authors.

## Results

Selected ICU stays are summarized in Table 2. ICU mortality rates differ between ICUs and are higher in Carevue stays. Seventy-fifth percentile ICU LOS also differ by location and are higher in Carevue stays. Other descriptive statistics of the MIMIC III data can be found in Dai et al [21].

Table 2: Selected ICU stays (Carevue 2001-2008; Metavision 2008-2012)

| | n | ICU LOS [IQR] (d) | ICU Mortality (%) |
|---|---|---|---|
| All | (28,817; 23,244) | ([0.75, 4.5]; [1.0, 3.75]) | (9.1; 7.7) |
| MICU | (10,399; 10,090) | ([1.1, 4.5]; [1.0,3.5]) | (11.6; 9.0) |
| SICU | (4,418; 4,277) | ([1.3, 5.5]; [1.0, 4.1]) | (10.1; 8.3) |
| CCU | (4,778; 2,731) | ([1.1, 4.1]; [1.1, 3.9]) | (9.3; 8.7) |
| CSRU | (5,799; 3,346) | ([1.1, 4.0]; [1.1,3.1]) | (3.5; 2.7) |
| TSICU | (3,423; 2,800) | ([1.0, 4.9]; [1.1; 4.1]) | (9.6; 7.0) |

For each observation, *vpres* was assigned the value (weight) that corresponds to the number of distinct vasopressors administered in the segment (Table 3). From Equation A, the estimated marginal effects (as odds ratios, OR) of ***OasisCompScores, vpresBins***, and *vpres* by ICU day are shown in Table 4.

Table 3: Vasopressor bins and weights

| n vasopressors (bins) | *vpres* (weights) |
|---|---|
| 0 | 0 |
| 1 | 4 |
| 2 | 8 |
| >=3 | 10 |

With the vasopressor bin weights in Table 3, and *vpres* set to the weight of its corresponding bin, *vpres* effects are approximately the right size, as designed. At the same time, ***vpresBins*** indicators are mostly insignificant, also as designed, suggesting that no additional information on vasopressor use remains in the ***vpresBins*** indicators. The *vpres* component score is very significant as expected since

vasopressors are important in critical care medicine. OASIS component scores for *heartrate, meanbp, resprate, temp, urineout, mechvent, gcs, and age* are frequently significant and are in the expected direction (OR>1). *PreICUlos* is significant only on the first two ICU days. *Electivesurgery* is not significant.

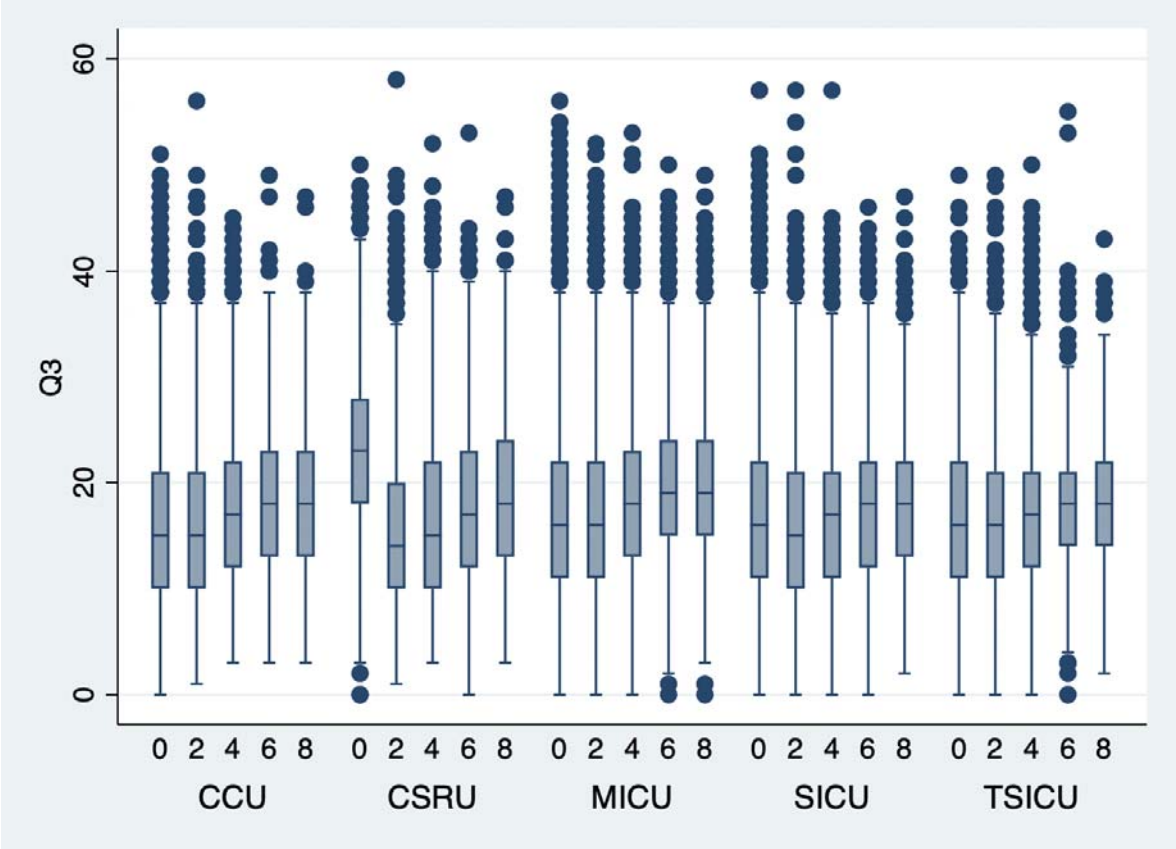
Table 4: Estimated effects (OR) of component scores on odds of ICU mortality (Eq A)

| | day0 | day1 | day2 | day3 | day4 | day5 | day6 | day7 | day8 | day9 |
|---|---|---|---|---|---|---|---|---|---|---|
| *age* | 1.16‡ | 1.11‡ | 1.12‡ | 1.14‡ | 1.15‡ | 1.16‡ | 1.16‡ | 1.18‡ | 1.18‡ | 1.19‡ |
| *preiculos* | 1.15‡ | 1.06** | 1.03 | 1.03 | 1.04 | 1.03 | 1.00 | 0.99 | 0.98 | 0.97 |
| *heartrate* | 1.30‡ | 1.24‡ | 1.22‡ | 1.17‡ | 1.17‡ | 1.18‡ | 1.18‡ | 1.16‡ | 1.13** | 1.19‡ |
| *meanbp* | 1.29‡ | 1.39‡ | 1.37‡ | 1.24‡ | 1.18** | 1.33‡ | 1.33‡ | 1.33‡ | 1.15* | 1.15 |
| *resprate* | 1.15‡ | 1.15‡ | 1.14‡ | 1.14‡ | 1.10‡ | 1.12‡ | 1.09** | 1.06* | 1.08** | 1.08* |
| *temp* | 1.22‡ | 1.14‡ | 1.10** | 1.17‡ | 1.17‡ | 1.04 | 1.08 | 1.07 | 1.15* | 1.10 |
| *urineout* | 1.14‡ | 1.12‡ | 1.14‡ | 1.12‡ | 1.12‡ | 1.13‡ | 1.12‡ | 1.12‡ | 1.11‡ | 1.10‡ |
| *mechvent* | 1.18‡ | 1.24‡ | 1.25‡ | 1.21‡ | 1.19‡ | 1.19‡ | 1.20‡ | 1.19‡ | 1.18‡ | 1.19‡ |
| *gcs* | 1.16‡ | 1.16‡ | 1.21‡ | 1.15‡ | 1.19‡ | 1.14‡ | 1.12‡ | 1.08* | 1.10* | 1.12** |
| *electivesurg* | 0.88 | 1.13 | 0.94 | 1.10 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 0.92 |
| | | | | | | | | | | |
| *vpres* | 1.10‡ | 1.11‡ | 1.12‡ | 1.11‡ | 1.18‡ | 1.16‡ | 1.14‡ | 1.17‡ | 1.18‡ | 1.21‡ |
| ***vpresBins*** | | | | | | | | | | |
| *0 (base)* | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| *1* | 1.01 | 1.15 | 1.20 | 1.41** | 1.17 | 1.07 | 1.22 | 1.16 | 1.04 | 0.92 |
| *2* | 0.79 | 0.89 | 1.02 | 1.35 | 0.85 | 0.96 | 1.43 | 0.88 | 0.99 | 0.84 |
| *>=3* | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |

* p<0.05, ** p<0.01, ‡p<0.001

Q3 was defined as the sum of 9 component scores: *vpres, heartrate, meanbp, resprate, temp, urineout, mechvent, gcs, and age*. *PreICUlos* and *Electivesurgery* were not used. Boxplots of Q3 by ICU across ICU days are shown in Fig 2. Median Q3 values are typically between 15 and 20 with interquartile ranges from 10 to 25. Based on these distributions, we selected Q3=5,15,25, and 35 to show Q3 marginal effects at different levels of Q3.

Figure 2: Q3 scores for the first segments of ICU days 0, 2, 4, 6, 8




**ICU mortality**

Parameter estimates for Equation B are shown in Table 5. They show that the response is quadratic in Q3 (concave down) through day3, then linear in Q3 from day4 through day9. The marginal effects of Q3 at Q3=5,15, 25, and 35 are included. Pearson-Windmeijer and Stukel tests for miscalibration were insignificant ($p>0.05$) for all daily models.

Table 5: Logistic regression estimates (OR) for ICU mortality (Eq. B)

| | day0 | day1 | day2 | day3 | day4 | day5 | day6 | day7 | day8 | day9 |
|---|---|---|---|---|---|---|---|---|---|---|
| *n* | 28,346 | 20,225 | 13,657 | 9,691 | 7,437 | 5,963 | 4,927 | 4,177 | 3,633 | 3,197 |
| | | | | | | | | | | |
| *Q3* | 1.224‡ | 1.256‡ | 1.234‡ | 1.236‡ | 1.171‡ | 1.194‡ | 1.160‡ | 1.117‡ | 1.127‡ | 1.173‡ |
| $Q3^2$ | .9987‡ | .9984‡ | .9989** | .9985** | .9997 | .9992 | .9998 | .9995 | 1.000 | .9993 |
| | | | | | | | | | | |
| *Q3 ME at Q3=* | | | | | | | | | | |
| *5* | 1.21‡ (.012) | 1.24‡ (.015) | 1.22‡ (.017) | 1.22‡ (.020) | 1.17‡ (.021) | 1.18‡ (.026) | 1.16‡ (.028) | 1.16‡ (.031) | 1.13‡ (.030) | 1.17‡ (.032) |
| *15* | 1.18‡ (.006) | 1.20‡ (.008) | 1.19‡ (.009) | 1.18‡ (.010) | 1.16‡ (.010) | 1.17‡ (.013) | 1.15‡ (.014) | 1.15‡ (.015) | 1.14‡ (.015) | 1.15‡ (.016) |
| *25* | 1.15‡ (.004) | 1.16‡ (.004) | 1.17‡ (.005) | 1.15‡ (.006) | 1.15‡ (.006) | 1.15‡ (.007) | 1.15‡ (.008) | 1.14‡ (.008) | 1.14‡ (.008) | 1.14‡ (.008) |
| *35* | 1.12‡ (.008) | 1.12‡ (.009) | 1.14‡ (.011) | 1.12‡ (.013) | 1.14‡ (.015) | 1.13‡ (0.17) | 1.15‡ (.019) | 1.13‡ (.020) | 1.15‡ (.021) | 1.12‡ (.020) |
| | | | | | | | | | | |
| *MICU* (base) | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| *SICU* | 0.84** | 0.89 | 0.89 | 0.81* | 0.73** | 0.74** | 0.69** | 0.67‡ | 0.64‡ | 0.67** |
| *CCU* | 0.90 | 0.85* | 0.89 | 0.84 | 0.86 | 0.89 | 0.83 | 0.92 | 0.85 | 0.86 |
| *CSRU* | 0.18‡ | 0.27‡ | 0.29‡ | 0.31‡ | 0.31‡ | 0.33‡ | 0.32‡ | 0.33‡ | 0.33‡ | 0.38‡ |
| *TSICU* | 0.85* | 0.81* | 0.80* | 0.73** | 0.68** | 0.64** | 0.57‡ | 0.46‡ | 0.46‡ | 0.47‡ |

ME = marginal effect; (SE) = standard error; * $p<0.05$, ** $p<0.01$, ‡ $p<0.001$

Discrimination and calibration tests show that models estimated on Carevue data perform well on Metavision data where they discriminate well (AUROC/c-stat typically > 0.75, Fig 3) and calibrate well (HL $p>0.05$, Table 6). Spiegelhalter's tests were insignificant ($p>0.05$) throughout, further suggesting reasonable discrimination and calibration.

Figure 3: Equation B discrimination testing

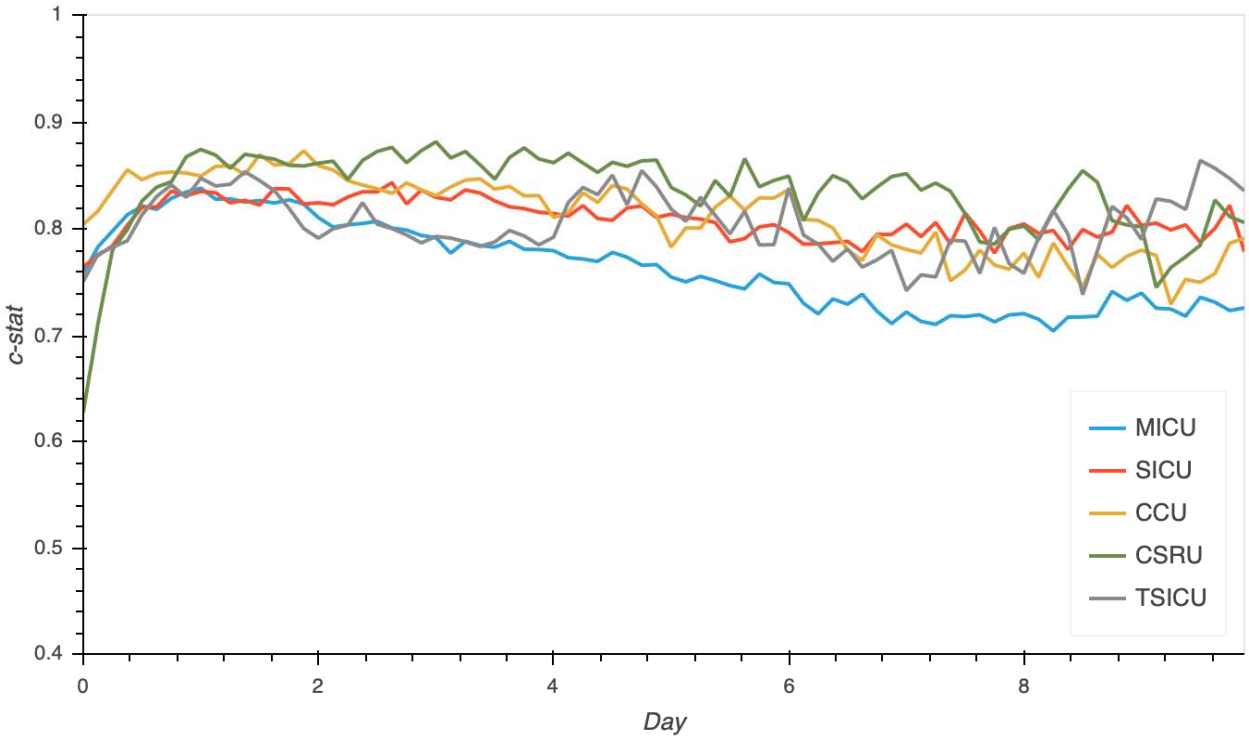



Table 6: Equation B calibration testing

| | Day | | | | |
|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 |
| MICU: | ++++++++ | ++++++++ | ++++++++ | ++++++++ | -++++++ |
| SICU: | ++++++++ | ++++++++ | ++++++-+- | ++++-+++ | ++++++++ |
| CCU: | ++++++++ | ++++++++ | ++++++++ | ++++++++ | ++++++++ |
| CSRU: | -++++++++ | ++++++++ | ++++++++ | ++++++++ | ++++++++ |
| TSICU: | +--++++++ | ++++++++ | ++++++++ | ++++++++ | ++++++++ |

| | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|
| MICU: | ++++++++ | +-++++++- | ---++++++ | +++++-++ | ++++++++ |
| SICU: | +-+++-++ | +++++++-+ | +-++++++ | ++++++++ | ++++++++ |
| CCU: | ++++++++ | ++++++++ | ++++++++ | ++++++++ | ++++++++ |
| CSRU: | ++++++++ | ++++++++ | ++++++++ | ++++++++ | ++++++++ |
| TSICU: | ++++++++ | ++++++++ | ++++++++ | ++++++++ | ++++++++ |

(+) calibrated at segment, HL p>0.05

The average marginal effect (AME) of Q3 on ICU mortality is the average change in probability caused by changing Q3 by one unit while leaving all other covariates unchanged. AME of Q3 by

location and day are shown in Table 7. The effects are on the probability scale and are mostly between 1 and 2 percentage points except for the CSRU where AME are smaller.

Table 7: Average marginal effects (probability scale) of Q3 on ICU mortality

| | day0 | day1 | day2 | day3 | day4 | day5 | day6 | day7 | day8 | day9 |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | |
| *MICU* | .012 (.000) | .013 (.000) | .015 (.001) | .017 (.001) | .018 (.001) | .019 (.001) | .020 (.001) | .020 (.001) | .020 (.001) | .020 (.001) |
| *SICU* | .011 (.000) | .012 (.000) | .014 (.001) | .015 (.001) | .015 (.001) | .016 (.001) | .017 (.001) | .016 (.001) | .016 (.001) | .017 (.001) |
| *CCU* | .011 (.000) | .012 (.001) | .014 (.001) | .015 (.001) | .017 (.001) | .018 (.001) | .018 (.001) | .019 (.001) | .019 (.001) | .019 (.002) |
| *CSRU* | .003 (.000) | .005 (.000) | .007 (.000) | .008 (.001) | .009 (.001) | .010 (.001) | .010 (.001) | .011 (.001) | .011 (.001) | .012 (.001) |
| *TSICU* | .011 (.000) | .011 (.001) | .013 (.001) | .014 (.001) | .015 (.001) | .015 (.001) | .015 (.001) | .013 (.001) | .013 (.001) | .013 (.001) |

(SE) = standard error; all estimates significant at $p<0.001$

**ICU Remaining LOS**

The marginal effects of Q3 on rLOS (Eq. C) are shown in Table 8, where for any estimated parameter *B*, the proportional effect of the corresponding variable is exp(*B*)-1. For small $|B|<0.25$, or so, this effect in percentage terms is approximately *B* times 100. For example, the day0 marginal effect of Q3 at Q3=15 is about 5.8% which means that a one-unit increase in Q3 from 15 to 16 increases the remaining ICU LOS by about 5.8%. Location indicators are frequently significant with longer rLOS in the SICU, CSRU, and TSICU than in the MICU and CCU. *Daypart* is sometimes significant, and *weekend* is significant only on day0, where rLOS is about 12% longer than non-weekend segments. The marginal effects of Q3 on rLOS at Q3 levels below 35 are always significant and positive. At Q3=35,

the marginal effect is sometimes positive, sometimes negative, and sometimes not significant. Most patients have Q3<25 (Fig 2).

Table 8: Poisson regression estimates for ICU rLOS (Eq. C)

| | day0 | day1 | day2 | day3 | day4 | day5 | day6 | day7 | day8 | day9 |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | |
| *Q3* | .090‡ | .143‡ | .163‡ | .178‡ | .167‡ | .159‡ | .180‡ | .142‡ | .142‡ | .174‡ |
| *Q3²* | -.001‡ | -.002‡ | -.002‡ | -.003‡ | -.003‡ | -.003‡ | -.003‡ | -.002‡ | -.002‡ | -.003‡ |
| | | | | | | | | | | |
| *Q3 ME at Q3=* | | | | | | | | | | |
| *5* | .079‡ (.006) | .125‡ (.007) | .139‡ (.009) | .151‡ (.014) | .140‡ (.011) | .132‡ (.015) | .148‡ (.014) | .118‡ (.020) | .120‡ (.020) | .142‡ (.019) |
| *15* | .058‡ (.003) | .090‡ (.003) | .090‡ (.004) | .096‡ (.007) | .086‡ (.005) | .078‡ (.007) | .085‡ (.006) | .072‡ (.007) | .076‡ (.008) | .079‡ (.009) |
| *25* | .037‡ (.003) | .054‡ (.003) | .041‡ (.003) | .042‡ (.004) | .031‡ (.005) | .023‡ (.005) | .022‡ (.005) | .025** (.009) | .032‡ (.007) | .015* (.006) |
| *35* | .016** (.006) | .018** (.007) | -.008 (.008) | -.013 (.010) | -.023* (.011) | -.031* (.013) | -.041** (.013) | -.021 (.022) | -.013 (.018) | -.048** (.014) |
| | | | | | | | | | | |
| *MICU* (base) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| *SICU* | 0.23‡ | 0.33‡ | 0.30‡ | 0.36‡ | 0.39‡ | 0.36‡ | 0.26‡ | 0.21** | 0.18 | 0.36‡ |
| *CCU* | 0.09 | -0.08 | 0.06 | -0.02 | -0.15 | 0.00 | -0.34‡ | 0.05 | -0.15 | -0.09 |
| *CSRU* | -0.24‡ | 0.00 | 0.09 | 0.29‡ | 0.08 | 0.25** | 0.09 | 0.19 | 0.36‡ | 0.29** |
| *TSICU* | 0.17** | 0.24‡ | 0.32‡ | 0.32‡ | 0.18* | 0.20** | 0.16 | 0.14 | 0.15 | 0.15 |
| | | | | | | | | | | |
| *daypart* | | | | | | | | | | |
| *0* (base) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| *1* | 0.13* | -0.08 | -0.05 | -0.13 | -0.02 | -0.07 | -0.07 | 0.01 | -0.24 | -0.10 |
| *2* | 0.03‡ | -0.12 | -0.24** | -0.07 | -0.10 | -0.05 | -0.15 | -0.26* | -0.04 | 0.15 |
| *3* | -0.08 | -0.11 | -0.18* | -0.09 | 0.09 | -0.10 | 0.05 | 0.01 | -0.05 | 0.04 |
| *4* | 0.09 | 0.07 | 0.08 | 0.06 | 0.10 | 0.10 | -0.01 | -0.12 | 0.13 | -0.08 |
| *5* | 0.05 | 0.09 | -0.05 | 0.07 | 0.07 | 0.03 | 0.14 | -0.11 | 0.08 | 0.01 |
| *weekend* | 0.12‡ | 0.02 | 0.06 | 0.05 | 0.03 | -0.04 | 0.04 | 0.05 | 0.08 | 0.03 |

ME = marginal effect; (SE) = standard error; * p<0.05, ** p<0.01, ‡ p<0.001

The average marginal effects of Q3 on rLOS (in hours) are shown in Table 9. They increase over the first few ICU days with most between 5 and 10 hours.

Table 9: Average marginal effects (in hours) of Q3 on rLOS

| | day0 | day1 | day2 | day3 | day4 | day5 | day6 | day7 | day8 | day9 |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | |
| *MICU* | 4.3 (.19) | 6.2 (.24) | 6.5 (.30) | 7.9 (.42) | 7.5 (.42) | 7.5 (.51) | 8.7 (.63) | 7.8 (.82) | 9.6 (.87) | 7.4 (.75) |
| *SICU* | 5.4 (.28) | 8.6 (.41) | 8.7 (.48) | 11.3 (.80) | 11.1 (.85) | 10.8 (.92) | 11.2 (.96) | 9.7 (1.1) | 11.5 (1.3) | 10.6 (1.4) |
| *CCU* | 4.7 (.27) | 5.7 (.27) | 6.9 (.47) | 7.7 (.64) | 6.5 (.57) | 7.5 (.85) | 6.2 (.61) | 8.2 (1.1) | 8.3 (1.4) | 6.8 (1.0) |
| *CSRU* | 3.4 (.19) | 6.2 (.31) | 7.1 (.53) | 10.5 (.78) | 8.2 (.73) | 9.6 (1.1) | 9.5 (.88) | 9.5 (1.4) | 13.8 (1.3) | 9.9 (1.2) |
| *TSICU* | 5.2 (.33) | 7.9 (.44) | 8.9 (.61) | 10.9 (.83) | 9.0 (.69) | 9.2 (.86) | 10.2 (1.1) | 9.0 (1.2) | 11.1 (1.3) | 8.6 (1.1) |

(SE) = standard error; all estimates significant at $p<0.001$

## Discussion

Acuity scores are used to calculate risk-adjusted mortality rates (actual vs. predicted) for benchmarking and quality improvement [22,23]. Existing ICU acuity scores are typically calculated only at the end of the first ICU day [1-7]. OASIS is non-proprietary and performs about as well as other ICU scores while using fewer variables. However, like other scores, it has only been tested at the end of the first ICU day [10]. Real-time models of ICU acuity use many clinical variables individually instead of one acuity score [8-10], which is limiting if a single score is needed as a dependent variable elsewhere.

In this study, we created Q3, an acuity score computed every 3 hours, and used it in tractable models of ICU mortality and ICU remaining LOS. Q3 is the sum of a new weighted vasopressor score and 8 of the 10 OASIS component scores. Logistic ICU mortality models quadratic in Q3 show no evidence of misspecification and discriminate well (AUROC ~ 0.72 – 0.85) and calibrate well (Hosemer-Lemeshow $p>0.05$, Spiegelhalter's z $p>0.05$) on newer ICU stays in all ICUs over the first 10 days of the ICU stay. The logistic ICU mortality model is quadratic in Q3 with the quadratic effect

different than OR=1 in the first 4 ICU days (Table 5). A one-unit increase in Q3 from a typical level of Q3=15 (Fig 2) affects the odds of ICU mortality by a factor of 1.14 to 1.20 ($p<0.001$, Table 5) and the ICU remaining LOS by 5.8 to 9.6% ($p<0.001$, Table 8), depending on the ICU day. The average effect of a one-unit change in Q3 (AME) on ICU mortality are often between 1 and 2 percentage points depending on location and ICU day (Table 7), and the AME of Q3 on ICU remaining LOS are usually between 5 and 10 hours depending on location and ICU day (Table 9).

The marginal effects of Q3 on rLOS are positive at Q3=5, 15, and 25, but at Q3=35, they are inconsistent (Table 8). Since Q3 increases the odds of mortality at all levels of Q3 (Table 5), mortality logically truncates rLOS, and increasing Q3 logically extends LOS not ending in death, the effects of Q3 on rLOS work through mortality and non-mortality causal paths, one plausibly increasing rLOS and the other shortening it. When they are roughly comparable, as they appear to be at high levels of Q3, the marginal effects of Q3 are sometimes positive, sometimes negative, and sometimes insignificant. Ad hoc methods to control for mortality in LOS models include restricting the model to only survivors, assigning large rLOS values to stays censored by death, or adding a mortality indicator. Each of these approaches, however, is problematic since it conditions the regression on a *future* mortality event [24]. Therefore, we chose to leave all stays in and interpret the effects plainly.

We think of Q3 as the residual acuity observable through Q3 components, while location indicators in our models account for average location-specific differences in comorbidities, case mix, and acuity unobserved by Q3. Without location indicators, the average unobserved acuity would be entirely contained in the constant terms of the equations, without location-based resolution. Location indicators work to capture differences between unobserved comorbidities between locations by comparing them to the base level location (MICU). This creates better calibrated models.

Although not shown, we conducted experiments of Q3 without the vasopressor score component. In those experiments, daily estimated models of Equation B show evidence of functional misspecification with several significant Stukel tests ($p<0.05$), and discriminate less well with c-statistics 1-2 points lower, typically, occasionally dropping below 0.7. This makes sense since *vpres*, the vasopressor score, is extremely significant in addition to other OASIS component scores (Table 4). For these reasons, Q3 with the vasopressor score component is a better score than without it.

To our knowledge, there are several other examples of real-time ICU acuity scores [8-10]. Szolovits uses the MIMIC II database and mortality models that depend on dozens of variables including labs and comorbidities [8]. Overall Day 3 AUROC is about 0.85, comparable to our results (Fig 2, day 2). Location-specific results were not given, and selected model components were not weighted and combined into a single acuity score, like OASIS and Q3. Shickel et. al. use 14 variables including select labs and medication use [9]. It achieves AUROC approaching 0.9 by 96 hours. However, like other deep learning models, its trained neural network is not interpretable. So, while it predicts well, the likely complex, non-linear relationships of input variables encoded by the neural network are currently undiscoverable. Johnson et al use the MIMIC III database to evaluate several real-time mortality prediction models [10]. Features extracted from about 40 clinical variables were computed over 4-hour intervals and were used to predict in-hospital mortality using several model types including logistic regression (LR) and gradient boosting decision trees. The gradient boosting model achieved a high AUROC of 0.93 with LR closely behind at 0.90. No data on model discrimination or calibration by ICU location or at different times during the ICU stay is given, and since many variables are used to capture acuity, no estimated marginal effects of a single acuity score can be given.

We believe that our work contributes to the literature on acuity scores and complements the work of the others, especially the OASIS research team. Q3 is a single score comprised of only 9 variables.

When used in tractable and well-behaved models of ICU mortality, Q3 significantly affects the odds of mortality in different ICUs across the first 10 days of the ICU stay (Tabe 5). The marginal effects of Q3 change by the level of Q3 and ICU day (Table 5), and the average marginal effects of Q3 on mortality and ICU remaining LOS are statistically significant and clinically meaningful (Table 7).

Future applications of Q3 include its use as a dependent variable in models that explore the effects of other time-changing variables on ICU acuity, and in estimating the aggregate acuity of subsets of patients over time to help stratify ICU patients in-stay, and to assist with ICU workload analysis.

## Limitations

The MIMIC database contains data from one hospital. However, the data are comprised of many years of observations from different ICUs and patient types.

## Conclusion

Q3 has significant effects on ICU mortality and ICU remaining LOS. Depending on location and ICU day, a one-unit increase in Q3 increases the probability of ICU mortality by or 1-2 percentage points ($p<0.001$) and ICU remaining LOS by 5 to 10 hours ($p<0.001$). ICU mortality models that use Q3 discriminate well and are calibrated in different ICUs across the first 10 days of the ICU stay. Given its meaningful effects and favorable performance characteristics, we believe that Q3 could be used as a dependent variable in explanatory models to help elucidate mechanisms of changing ICU acuity.

## Acknowledgements

None

## Competing Interests

All authors have a financial interest in Indicator Sciences, LLC

## Funding

Indicator Sciences, LLC

## References

[1] W.A Knaus, D.P. Wagner, E.A. Draper, J.E. Zimmerman, M. Bergner, P.G. Bastos, C.A. Sirio, D.J. Murphy, T. Lotring, A. Damiano, et al. The APACHE III prognostic system: Risk prediction of hospital mortality for critically ill hospitalized adults, Chest, vol. 100, no. 6, pp. 1619-1636, 1991. doi: 10.1378/chest.100.6.1619.

[2] J.R LeGall, S. Lemeshow, and F. Saulnier, A new simplified acute physiology score (SAPS-II) based on a European North American multicenter study, JAMA, vol. 270, pp. 2957-2963, Dec 22 1993. doi: 10.1001/jama.270.24.2957.

[3] S. Lemeshow, D. Teres, and J. Klar, Mortality probability model (MPM II) based on an international cohort of intensive care unit patients, JAMA, vol. 270, pp. 2478-2486, 1993.

[4] J.L. Vincent, R Moreno, J. Takala, S. Willats, A. De Mendoca, H. Bruining, C.K. Reinhart, P.M. Suter, and L.G. Thijs. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. *Intensive Care Medicine*, 22:707–710, 1996. doi: 10.1007/BF01709751.

[5] J.L. Vincent and R. Moreno, Clinical review: scoring systems in the critically ill, Critical care, vol. 14, p. 207, Jan. 2010. doi: 10.1186/cc8204.

[6] M.T. Keegan, G. Ognjen, and A. Bekele. Severity of illness scoring systems in the intensive care unit. *Critical care medicine*, 39(1):163–169, 2011.

[7] A.E. Johnson, A.A. Kramer, and G.D Clifford. A new severity of illness scale using a subset of Acute Physiology and Chronic Health Evaluation data elements shows comparable predictive accuracy. Crit Care Med. 2013 Jul;41(7):1711-8. doi: 10.1097/CCM.0b013e31828a24fe

[8] C.W. Hug and P. Szolovits. ICU acuity: real-time models versus daily models. *AMIA Annu Symp Proc*. 2009; 2009:260-264. Published 2009 Nov 14. doi: 10.1097/CCM.0b013e3181f96f81.

[9] B. Shickel, T.J. Loftus, L. Adhikari, *et al.* DeepSOFA: A Continuous Acuity Score for Critically Ill Patients using Clinically Interpretable Deep Learning. *Sci Rep* 9, 1879 (2019). doi: 10.1038/s41598-019-38491-0.

[10] A.E. Johnson and R.G. Mark. Real-time mortality prediction in the Intensive Care Unit. *AMIA Annu Symp Proc*. 2018; 2017:994-1003.

[11] A.E. Johnson, T.J. Pollard, L Shen, L. Lehman, et al. MIMIC-III, a freely accessible critical care database. *Scientific Data* 2016;3(1):160035. doi: 10.1038/sdata.2016.35.

[12] A.E. Johnson, T.J Pollard, and R.G Mark (2016). MIMIC-III Clinical Database (version 1.4). *PhysioNet*. doi: 10.13026/C2XW26.

[13] F. Windmeijer. (1990) The asymptotic distribution of the sum of weighted squared residuals in binary choice models. Statistica Neerlandica 44, 2: 69–78. doi: 10.1111/j.1467-9574.1990.tb01527.x.

[14] T.A. Stukel. (1988) Generalized logistic models. Journal of the American Statistical Association 83: 426–431. doi: 10.1080/01621459.1988.10478613

[15] D.W. Hosemer, S. Lemeshow, and R.X. Sturdivant. Assessing the Fit of the Model: Applied Logistic Regression. 3rd ed. Wiley-Blackwell; 2013:167-168. doi: 10.1002/0471722146.ch5

[16] P. Paul, M.L Pennell, and S. Lemeshow. Standardizing the power of the Hosmer-Lemeshow goodness of fit test in large data sets. Statistics in Medicine. Jan 2013; 32:67-80. doi: 10.1002/sim.5525.

[17] ICU Outcomes (Mortality and Length of Stay) Methods, Data Collection Tool and Data. Philip R. Lee Institute for Health Policy Studies. Retrieved Mar 4, 2022 from https://healthpolicy.ucsf.edu/icu-outcomes.

[18] A.S. Gomes,M.M Kluck, J. Riboldi, and J.M. Fachel. Mortality prediction model using data from the Hospital Information System. Revista de Saude Publica. Oct 2010; 44(5):934-41. doi: 10.1590/S0034-89102010005000037.

[19] G.W. Brier. Verification of forecasts expressed in terms of probability. Mon Weather Rev. 1950;78(1):1–3. doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

[20] J.M. Wooldridge. Econometric Analysis of Cross Section and Panel Data, MIT Press Books, The MIT Press, edition 2, volume 1, number 0262232588, Dec 2010. Chapter 18

[21] Z. Dai, S. Liu, J. Wu, M. Li, J. Liu, and K. Li. (2020) Analysis of adult disease characteristics and mortality on MIMIC-III. PLoS ONE 15(4): e0232176. doi: 10.1371/journal. pone.0232176.

[22] T.L. Higgins. Quantifying risk and benchmarking performance in the adult intensive care unit. J Intensive Care Med. 2007 May-Jun;22(3):141-56. doi: 10.1177/0885066607299520.

[23] M.J. Breslow and O. Badawi. Severity scoring in the critically ill: part 2: maximizing value from outcome prediction scoring systems. Chest. 2012 Feb;141(2):518-527. doi: 10.1378/chest.11-0331.

[24] G.N. Brock, C. Barnes, J.A. Ramirez, *et al.* How to handle mortality when investigating length of hospital stay and time to clinical stability. *BMC Med Res Methodol* 11, 144 (2011). doi: 10.1186/1471-2288-11-144.