Cureus
Part of SPRINGER NATURE

Open Access Original Article

Treatment Limitation and Risk-Adjusted Mortality Benchmarking in 190 Intensive Care Units in the United States

Review began 10/10/2026
Review ended 11/10/2026
Published 11/12/2026

© **Copyright** 2026
Nyambod et al. This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

DOI: 10.7759/cureus.119580

Fabrice T. Nyambod [1], Aja Wagoh [2], Murishi Onesphore [3], January G. Msemakweli [4]

1. Department of International Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, USA 2. Department of Nursing, University of Buea, Buea, CMR 3. Department of Public Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, USA 4. Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, USA

Corresponding author: Fabrice T. Nyambod, fabricetiku1@gmail.com

Abstract

Introduction: Intensive care units (ICUs) are often ranked using standardized mortality ratios derived from the Acute Physiology and Chronic Health Evaluation (APACHE) IVa model, which does not include a term for treatment limitation. We assessed three primary questions: (1) whether APACHE IVa calibration depended on documented treatment limitation, (2) whether between-unit variation in treatment limitation persisted after balancing measured case mix, and (3) whether accounting for treatment limitation changed unit rankings.

Methods: This was a retrospective cohort study of 136,236 adult admissions to 190 US ICUs included in the eICU Collaborative Research Database (2014-2015). The most restrictive treatment limitation documented within 24 hours and its timing were classified into categories, ranging from full therapy to comfort measures only. Units in the highest and lowest quintiles of documented treatment limitation were compared before and after inverse probability weighting based on 12 characteristics. Units were ranked according to their standardized mortality ratios, with and without modeled treatment limitation.

Results: Treatment limitation was documented in 13,496 (9.9%) of 136,236 admissions. The standardized mortality ratio was 0.62 (95% confidence interval: 0.60-0.63) under full therapy and 1.87 (1.73-2.02) under comfort measures, regardless of whether the limitation was documented within four hours (2.03) or later (1.84). The highest quintile documented treatment limitation 3.58 times as often as the lowest quintile before weighting and 2.89 times as often after weighting. Modeling treatment limitation changed the quartile assignment of 38 (23.5%) of 162 units and the outlier status of 13 (8.0%). Excluding the 878 (0.6%) admissions involving comfort measures reclassified 18 (11.1%) units.

Conclusions: APACHE IVa calibration was poorer as treatment limitation became more restrictive. Between-unit variation in documented treatment limitation persisted after balancing measured case mix, and accounting for treatment limitation changed the ranking quartile of 38 (23.5%) of 162 units. Standardized mortality ratios may partly reflect differences in treatment limitation as well as quality of care.

Categories: Palliative Care, Quality Improvement, Epidemiology/Public Health
Keywords: benchmarking, hospital mortality, intensive care units, quality indicators, resuscitation orders, risk adjustment, severity of illness index

Introduction

End-of-life practices vary widely between intensive care units (ICUs). Withdrawal of life-sustaining treatment among patients who die ranges from 0% to 84% across published cohorts [1], with marked regional variation reported in a study of 199 units across 36 countries [2], as well as variation between units within individual countries [1]. Two academic centers at opposite extremes of treatment intensity differed in their norms for decision-making regarding life-sustaining treatment [3]. Hart et al. studied 277,693 admissions to 141 US units and found that the proportion of patients admitted with existing treatment limitations ranged from under 1% to 20.9% after risk adjustment, a variation that could not be explained by measurable center characteristics [4].

Treatment limitation is common in critical illness; do-not-resuscitate orders or orders limiting life-sustaining treatment were recorded for 211 (36.6%) patients in one medical intensive care cohort [5]. In another cohort of 200 intensive care decedents, 59 (29.5%) had a do-not-resuscitate order placed within 48 hours of admission [6]. Mortality among patients with such orders is substantially higher than that among patients without them [5], and the presence of an order has been associated with fewer operations, but not fewer other invasive interventions [7]. These orders are not homogeneous. A limitation documented at admission may reflect a preference expressed before critical illness or a prognosis evident at presentation. A

limitation documented after a period of full treatment may reflect the outcome of a time-limited trial of intensive care [8], the arrival of a family member, or the communication practices of the clinician or unit, and may partly reflect the care delivered.

Severity-of-illness scores estimate the probability of hospital death using measures intended to be independent of therapy [9] and thus implicitly assume full treatment. The Acute Physiology and Chronic Health Evaluation (APACHE) IV model was developed and validated using 110,558 admissions to 104 units, and its covariates describe first-day physiology, age, comorbidity, and admission diagnosis [10]. APACHE IVa does not include a covariate for a decision to withhold resuscitation; therefore, an order not to resuscitate alters the probability of death without changing any variable observed by the model. In contrast, the Mortality Probability Admission Model (MPM0-III) includes resuscitation status at admission [11]. The calibration of these models deteriorates as clinical practice changes and differs between populations, generally failing before discrimination does [12,13].

Standardized mortality ratios derived from these scores are the principal measures used in intensive care benchmarking. Units are ranked and, in some systems, reimbursed based on the assumption that the ratio isolates quality of care once case mix is accounted for [14,15]. Several studies have questioned this assumption. The choice of scoring system determines which units are classified as outliers [14,15], while model misspecification and modeling strategy affect unit-level estimates [16]. Accounting for do-not-resuscitate status changed hospital mortality rankings for pneumonia [17], and intensive care units less inclined to limit treatment had higher standardized mortality ratios [18]. Changes in documented care limitations have also been associated with incorrect APACHE IVa predictions [19]. The present study builds on these reports by examining, within a single multicenter cohort, how documented treatment limitation relates to APACHE IVa calibration, whether between-unit differences in treatment limitation persist after balancing measured case mix, and how accounting for treatment limitation changes the ranking of intensive care units.

We examined 136,236 admissions to 190 US intensive care units with three primary objectives: (1) to assess whether APACHE IVa calibration depended on documented treatment limitation, (2) to assess whether between-unit variation in treatment limitation persisted after balancing measured case mix, and (3) to assess whether accounting for treatment limitation changed unit rankings. A secondary, exploratory objective, added after completion of the primary analyses, was to assess whether these findings remained consistent when only orders written within four hours of admission were counted. The study describes the consequences of treatment limitation for benchmarking; it does not determine whether benchmarking models should adjust for it.

Materials And Methods

Study design and data source

This analysis used only publicly available data whose deidentification had been certified as meeting United States Safe Harbor standards; therefore, research ethics board approval and informed consent were not required. This retrospective cohort study was reported in accordance with the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) recommendations [20] and, for the assessment of APACHE IVa, the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) recommendations [21]. We analyzed version 2.0 of the eICU Collaborative Research Database, which contains 200,859 unit stays from 208 US hospitals, ranging from small community hospitals to large academic centers, that participated in a tele-critical care program in 2014 and 2015 [22]. The database was accessed under the PhysioNet credentialed data use agreement.

Cohort

Of the 200,859 adult unit stays, 52,327 (26.1%) had no APACHE IVa result, partly because the scoring rules exclude readmissions, admissions from another unit, burns, and stays shorter than four hours [10,15]. Additionally, 12,296 (6.1%) had no usable predicted mortality, leaving 136,236 (67.8%) stays across 190 units. Unit-level analyses were restricted to the 162 (85.3%) units contributing 100 or more stays (135,325 admissions; 99.3% of the cohort). The unit of analysis was the unit stay, with standard errors clustered by unit.

Exposure and timing

Treatment limitation was identified from structured care plan documentation (entries in the care limitation group of the carePlanGeneral table, referred to here as code status documentation), which records a care limitation value and its corresponding timestamp. In this article, treatment limitation refers to any documented order restricting life-sustaining treatment, ranging from resuscitation limitation to comfort measures only. For each stay, we recorded the most restrictive value documented within 24 hours of admission, corresponding to the period over which APACHE IVa collects its physiological inputs [10]. Full therapy served as the reference category. Resuscitation limitation comprised do-not-resuscitate, no cardiopulmonary resuscitation, no intubation, and no cardioversion; partial withdrawal comprised no vasopressors or inotropes, no augmentation of care, no blood products, and no blood draws; and comfort

measures only represented the most restrictive category. Stays with no entry were reported but not interpreted.

Each stay with a treatment limitation was classified according to the timing of the first order at its most restrictive tier: early, defined as within four hours of admission, including orders documented before admission; or late, defined as between 4 and 24 hours after admission. The four-hour threshold falls within the admission assessment period and precedes any sustained trial of treatment; one hour was also examined as a stricter cutoff. A record-based definition identified orders preceded by a full therapy entry, indicating de-escalation after an initial decision to treat. The timing analyses were added after completion of the primary analyses and were not prespecified.

Outcome and statistical analysis

The outcome was in-hospital mortality, which was compared with APACHE IVa-predicted hospital mortality. Calibration was assessed through logistic recalibration of observed mortality against the logit of predicted mortality and, within strata, using the standardized mortality ratio with exact Poisson 95% confidence intervals (CIs) and the integrated calibration index [23]. The independent contribution of limitation status was estimated by adding it to a recalibration model containing the APACHE linear predictor, with malignancy, which has been associated with heterogeneous APACHE IV calibration [24], included as a comparator. Limitation status was entered into these models as a categorical term, with full therapy as the reference category, while admissions with no documented entry formed a separate category. Discrimination was assessed using a random 30% holdout sample, based on the area under the receiver operating characteristic curve and the Brier score.

To assess whether between-unit differences persisted after balancing measured case mix, we compared units in the highest and lowest quintiles of limitation rate before and after inverse probability of treatment weighting based on 12 patient characteristics: age, sex, APACHE IVa score, acute physiology score, predicted mortality, metastatic or hematologic cancer, immunosuppression, cirrhosis, hepatic failure, diabetes, mechanical ventilation on day 1, and admission from the emergency department. These variables were selected because they comprise the inputs and outputs of APACHE IVa, comorbidities associated with both prognosis and treatment limitation, first-day markers of acuity and care pathway, and sex. Propensity scores for membership in the highest quintile were estimated using logistic regression, and the weights targeted the average treatment effect. Balance was assessed using a standardized mean difference threshold of 0.1 [25]. If measured case mix accounted for the difference, balancing it should eliminate that difference. Between-unit variation was also summarized using the median odds ratio derived from mixed-effects models with a unit-level random intercept [26].

For benchmarking, we calculated expected deaths per unit using a recalibrated APACHE IVa model, both with and without limitation status (logistic regression of in-hospital death on the APACHE IVa linear predictor, with or without limitation status included as a categorical term). Units were ranked according to their standardized mortality ratios under each model, and the rankings were compared in terms of rank shifts, quartile changes, and outlier status against exact Poisson limits [27]. We also evaluated the exclusion of admissions involving comfort measures and the exclusion of all admissions with any treatment limitation, refitting the model for each retained cohort. Each analysis was repeated according to the timing of the order, including a ranking analysis that counted only early orders, with late orders recoded as full therapy. In a sensitivity analysis, admissions with no documented code status were recoded as full therapy. Age, censored above 89 years, was set to 91; no imputation was performed. All analyses were conducted using R version 4.6 (R Foundation for Statistical Computing, Vienna, Austria), with the data.table, lme4, sandwich, WeightIt, and cobalt packages.

Results

Cohort

Among 136,236 admissions, mean APACHE IVa predicted mortality was 11.8%, and in-hospital death occurred in 12,061 (8.9%) admissions, with an overall standardized mortality ratio of 0.75. Code status was documented within 24 hours in 130,873 (96.1%) admissions. A limitation was present in 13,496 admissions (9.9%): a resuscitation limitation in 11,929 (8.8%), partial withdrawal in 689 (0.5%), and comfort measures only in 878 (0.6%). Patients with any limitation were older and more severely ill (Table *1*).

Characteristic	Full therapy	Resuscitation limitation	Partial withdrawal	Comfort measures	Not documented
Stays, N (%)	117,377 (86.2%)	11,929 (8.8%)	689 (0.5%)	878 (0.6%)	5,363 (3.9%)
Age, median years	63.0	77.0	76.0	74.0	64.0
Age over 89, N (%)	2,552 (2.2%)	1,863 (15.6%)	131 (19.0%)	121 (13.8%)	207 (3.9%)
Female, N (%)	52,919 (45.1%)	6,449 (54.1%)	385 (55.9%)	443 (50.5%)	2,411 (45.0%)
APACHE IVa score, median	49.0	66.0	69.0	85.0	50.0
Predicted mortality, %	10.5	22.3	26.2	39.6	11.2
Deaths, N (%)	7,605 (6.5%)	3,115 (26.1%)	220 (31.9%)	651 (74.1%)	470 (8.8%)
Standardized mortality ratio	0.62	1.17	1.22	1.87	0.78
Malignancy, N (%)	3,389 (2.9%)	729 (6.1%)	48 (7.0%)	60 (6.8%)	100 (1.9%)
Immunosuppression, N (%)	2,983 (2.5%)	514 (4.3%)	19 (2.8%)	53 (6.0%)	85 (1.6%)
Ventilated on day 1, N (%)	27,751 (23.6%)	3,492 (29.3%)	224 (32.5%)	402 (45.8%)	1,034 (19.3%)
Unit stay, median hours	43.2	43.1	36.3	18.2	43.1

TABLE 1: Characteristics of 136,236 intensive care admissions by treatment limitation status within 24 hours

Malignancy denotes a metastatic solid tumor or hematologic malignancy. Predicted mortality is the mean APACHE IVa-predicted hospital mortality. Data are from the eICU Collaborative Research Database [22].

APACHE: Acute Physiology and Chronic Health Evaluation.

APACHE IVa calibration by treatment limitation status

The overall calibration slope was 1.005, with an intercept of -0.425; the slope was near 1, but the model overpredicted death across the entire risk range. Calibration differed by limitation status (Table *2*, Figure *1A*). The standardized mortality ratio was 0.62 (95% CI, 0.60-0.63) under full therapy, 1.17 with resuscitation limitation, 1.22 with partial withdrawal, and 1.87 (95% CI, 1.73-2.02) under comfort measures only. The integrated calibration index was 0.34 under comfort measures compared with 0.04 to 0.06 in the other documented strata.

Code status	Admissions, N (%)	Predicted mortality, %	Deaths, N (%)	SMR	ICI	Emax
Full therapy	117,377 (86.2%)	10.5	7,605 (6.5%)	0.62	0.039	0.241
Resuscitation limitation	11,929 (8.8%)	22.3	3,115 (26.1%)	1.17	0.037	0.048
Partial withdrawal	689 (0.5%)	26.2	220 (31.9%)	1.22	0.057	0.078
Comfort measures	878 (0.6%)	39.6	651 (74.1%)	1.87	0.342	0.487
Not documented	5,363 (3.9%)	11.2	470 (8.8%)	0.78	0.022	0.125

TABLE 2: APACHE IVa calibration by treatment limitation status

Predicted mortality is the mean APACHE IVa-predicted hospital mortality. Data are from the eICU Collaborative Research Database [22].

APACHE: Acute Physiology and Chronic Health Evaluation, SMR: standardized mortality ratio, ICI: integrated calibration index, Emax: maximum absolute calibration error.

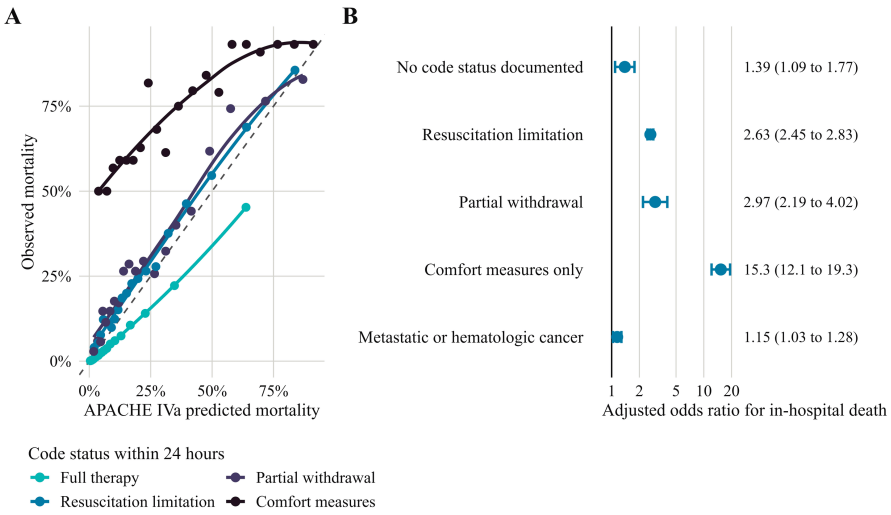


FIGURE 1: APACHE IVa performance by treatment limitation status

(A) Calibration curves. Points are 5% risk bins, and the dashed line indicates perfect calibration. (B) Adjusted odds ratios for in-hospital death from a logistic recalibration model, with full therapy as reference and standard errors clustered by unit.

Data are from the eICU Collaborative Research Database [22].

APACHE: Acute Physiology and Chronic Health Evaluation.

Image credits: This figure was created with R version 4.6 (R Foundation for Statistical Computing, Vienna, Austria) using the ggplot2 and patchwork packages.

Relative to full therapy, the adjusted odds of death were 2.63 with resuscitation limitation, 2.97 with partial withdrawal, and 15.3 under comfort measures only, compared with 1.15 for malignancy (Figure *1B*). Adding limitation status reduced model deviance by 3.6% against 0.0% for malignancy. On held-out data, it improved the area under the curve from 0.866 to 0.876 (DeLong $p < 0.001$) and reduced the Brier score from 0.062 to 0.058.

Variation between units after balancing measured case mix

Across 162 units, limitation rates ranged from 0 to 27.5% with a median of 10.5% (Figure *2A*). Before weighting, units in the highest limitation quintile differed from those in the lowest on two of 12 measured characteristics: age (standardized mean difference 0.197) and admission from the emergency department (0.304). Severity was nearly identical (-0.006 for the APACHE IVa score). After weighting, all 12 covariates were balanced within the 0.1 threshold (Figure *2B*). Despite this balance, the highest quintile documented limitation in 16.5% of admissions compared with 5.7% in the lowest, a ratio of 2.89, compared with 3.58 unweighted. About three-quarters of the excess therefore remained after measured case mix was balanced. In mixed-effects models, the median odds ratio rose from 1.69 to 1.80 after adjustment for case mix.

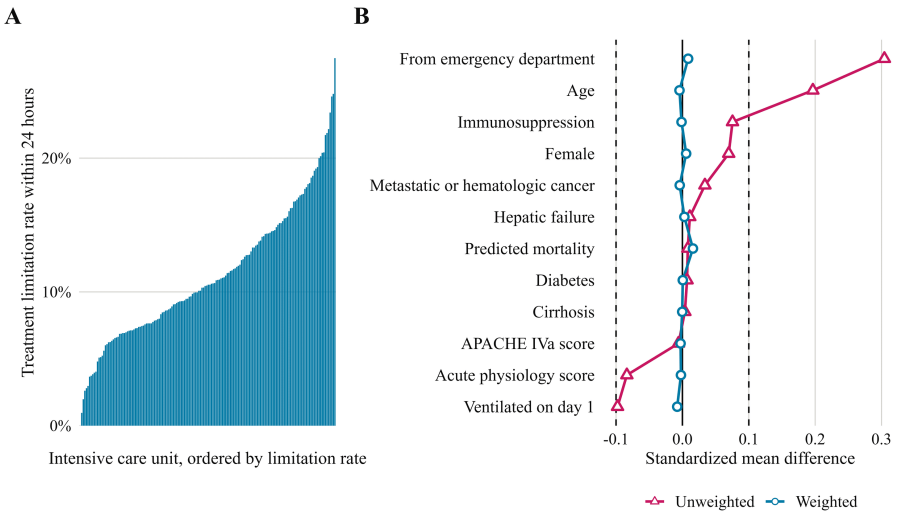


FIGURE 2: Between-unit variation in treatment limitation

(A) Limitation rate within 24 hours across 162 units with 100 or more admissions, ordered. (B) Standardized mean differences in patient characteristics between units in the highest and lowest limitation quintile, before and after inverse probability weighting.

Data are from the eICU Collaborative Research Database [22].

APACHE: Acute Physiology and Chronic Health Evaluation.

Image credits: This figure was created with R version 4.6 (R Foundation for Statistical Computing, Vienna, Austria) using the ggplot2, patchwork and cobalt packages.

Consequences for benchmarking

Adding limitation status preserved the overall ordering of units (Spearman correlation 0.949), although individual units moved, with a median absolute rank shift of 7.5 out of 162 positions. Thirty-eight units (23.5%) changed quartile, and 13 (8.0%) changed outlier classification; six (3.7%) outliers became unremarkable, and seven (4.3%) unremarkable units became outliers, three (1.9%) with high and four (2.5%) with low mortality (Figure *3B*). The direction of movement was related to each unit's limitation rate. The signed rank shift correlated with a unit's limitation rate at a Spearman coefficient of -0.689 (Figure *3A*). Units in the lowest limitation quintile moved from a standardized mortality ratio of 0.984 to 1.027 and those in the highest from 1.013 to 0.876.

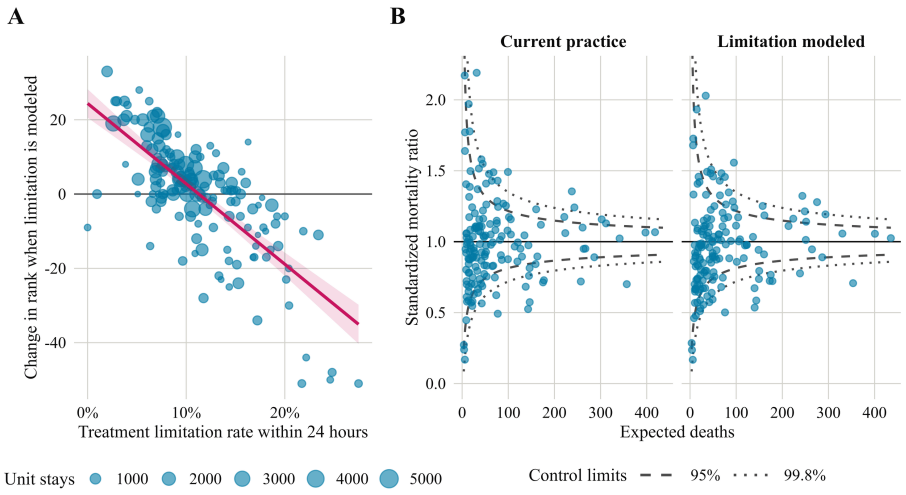


FIGURE 3: Effect of modeling treatment limitation on unit benchmarking

(A) Change in unit rank against limitation rate. Negative values indicate improvement. (B) Funnel plots of standardized mortality ratio against expected deaths under current practice and with limitation modeled, with 95% and 99.8% exact Poisson control limits.

Data are from the eICU Collaborative Research Database [22].

Image credits: This figure was created with R version 4.6 (R Foundation for Statistical Computing, Vienna, Austria) using the ggplot2 and patchwork packages.

Excluding the 878 (0.6%) admissions under comfort measures orders reclassified 18 units (11.1%) across quartiles, roughly half the effect of full adjustment. Excluding all 13,496 (9.9%) admissions with any limitation reclassified 46 units (28.4%) (Table 3). When the 5,363 (3.9%) admissions with no documented code status were treated as full therapy, the standardized mortality ratio under full therapy was 0.63 (95% CI 0.61-0.64), 36 (22.2%) units changed quartile, and 14 (8.6%) changed outlier classification.

Strategy	Admissions affected	Units changing quartile	Correlation with current
Current practice, APACHE IVa alone	0	Reference	1.000
Adjust for limitation status	0	38 (23.5%)	0.949
Exclude comfort measures only	878 (0.6%)	18 (11.1%)	0.986
Exclude any limitation	13,496 (9.9%)	46 (28.4%)	0.878

TABLE 3: Comparison of remediation strategies

Correlation is the Spearman rank correlation of standardized mortality ratios against current practice. Data are from the eICU Collaborative Research Database [22].

APACHE: Acute Physiology and Chronic Health Evaluation.

Timing of limitation orders

Comfort measures orders were documented a median of 666 minutes after unit admission (interquartile range 288-1052), with 187 (21.3%) within four hours, and resuscitation limitations at a median of 120 minutes. A full therapy entry preceded 547 (62.3%) comfort measures orders. Calibration failed at both time points. Among comfort measures admissions, the standardized mortality ratio was 2.03 (95% CI 1.69-2.43) for early and 1.84 (95% CI 1.69-2.00) for late orders, and the adjusted odds of death (12.9 and 16.2) did not differ (p = 0.28). Results were similar with a one-hour cut (standardized mortality ratio 2.02 against 1.87) and by whether a full therapy entry preceded the order (1.82 when preceded and 1.99 when not). For resuscitation limitation, late orders carried higher adjusted odds than early ones (3.25 against 2.31, p <

0.001).

Between-unit variation was larger for early orders. After weighting, units in the highest limitation quintile documented early limitation 3.53 times as often as those in the lowest (4.48 before weighting), compared with 2.20 for late orders. Counting only early orders, with later orders treated as full therapy, changed quartile assignment for 16 units (9.9%) and outlier classification for 6 (3.7%). Rank shift correlated with a unit's early limitation rate at a Spearman coefficient of -0.698. Excluding only comfort measures orders written within four hours of admission changed quartile assignment for six (3.7%) units and the outlier classification of none.

Discussion

This retrospective cohort study of 136,236 admissions to 190 US ICUs found that APACHE IVa calibration differed by documented treatment limitation, which the model does not include. The standardized mortality ratio was 0.62 among patients receiving full therapy and 1.87 among patients under comfort measures. Documentation of limitation varied severalfold between units, and the difference persisted after measured case mix was balanced. When limitation was added to the model, about one unit in four moved to a different ranking quartile. Each finding held when only orders written within four hours of admission were counted, and the between-unit variation was largest for those early orders.

Timing of the order

A comfort measures order within 24 hours may be written at admission or after a full initial effort, and most in this cohort followed a documented decision to treat fully. This heterogeneity did not change the main findings. The model was equally uninformative for early and late comfort measures orders, as would be expected if the order records a decision rather than a physiological state. Later resuscitation limitations were associated with higher mortality than earlier ones, possibly because they more often follow deterioration that the score captures only in part. Counting only early orders was the most conservative check against adjusting for a consequence of care. It reduced the benchmarking effect by more than half but did not remove it. Early orders least reflect care delivered in the unit, and the between-unit difference was largest for them. This suggests that the variation may reflect how units approach goals of care at admission more than how their patients fare during the first day.

Interpreting between-unit variation

These results extend Hart et al., who reported comparable variation after regression adjustment in a different database a decade earlier [4], and agree with the higher standardized mortality ratios Bertolini et al. found in units less inclined to limit treatment [18]. Our design equalized the covariate distributions rather than adjusting for them [25]. One possible explanation for the variation is how palliative care services are organized. Early palliative care consultation more than doubled transition to limitation orders in a randomized comparison [28]. The eICU database carries no information on palliative care staffing or goals-of-care protocols, so this explanation should be viewed as speculative. Because only measured characteristics were balanced, the remaining variation may also reflect unmeasured severity, patient preferences, or differences in how limitation is documented.

Consequences for benchmarking

Kramer et al. found that two risk models agreed on the direction and significance of the standardized mortality ratio in only 45% of units [15]. In this study, adding one variable already recorded in 130,873 (96.1%) charts changed the quartile assignment of about a quarter of units. This study does not resolve whether benchmarking models should adjust for limitation status. Adjustment treats limitation as patient preference and therefore as case mix; withholding adjustment treats it as modifiable unit behavior. An order written after an initial trial of treatment may in part be a product of that treatment, and adjusting for it risks crediting a unit for outcomes its own care produced. The same dilemma has been described for social risk factors [29], and misspecification of the adjustment model is itself a recognized source of bias [16].

Code status is a routine element of the admission record but enters neither the APACHE IVa model nor its exclusion rules. Benchmarking programs already exclude categories of patients, and the National Quality Forum model excludes 27.9% of admissions against 10.6% for APACHE IVa [15]. Options include excluding comfort measures orders written within four hours of admission, which affected one or two patients in a thousand and still reclassified six (3.7%) units. Alternatives are excluding all comfort measures admissions, six in a thousand, which recovered about half the effect, or reporting limitation rates alongside standardized mortality ratios. None of these requires new data collection. MPM0-III already includes resuscitation status at admission [11], which suggests that such adjustment is feasible within existing models.

Limitations

This study has limitations. First, treatment limitation is partly a consequence of prognosis, including prognostic information that APACHE IVa does not capture, a dependency that may bias mortality-based estimates [30]. Its association with mortality and the differences between units may therefore partly reflect

unmeasured severity rather than practice, and the findings should not be interpreted causally. Weighting balanced measured characteristics only [25], and cancer ascertainment relied on APACHE comorbidity flags. Second, we observed documentation and not practice. Documentation practices may differ between institutions, so part of the between-unit variation may reflect how limitation is recorded rather than how often it is decided, and delays in charting may misclassify some early orders as late. Third, the data describe 2014 and 2015 and derive from units in a single tele-critical care program [22]. Both the use and the documentation of treatment limitation may have changed since then and may differ in units outside this network. Fourth, the timing analyses bound rather than eliminate the hazard of mediator adjustment. They were not prespecified and should be read as exploratory.

Conclusions

APACHE IVa calibration was poorer as documented treatment limitation became more restrictive, and units differed severalfold in documenting limitation, a difference that persisted after measured case mix was balanced. Accounting for limitation changed the ranking of many units, even when only orders written within four hours of admission were counted. Risk-adjusted mortality comparisons may therefore partly reflect differences in treatment limitation as well as quality of care. Options that require no new data include excluding admissions under comfort measures orders and reporting unit limitation rates alongside the standardized mortality ratio.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Fabrice T. Nyambod, Murishi Onesphore, Aja Wagoh, January G. Msemakweli

Acquisition, analysis, or interpretation of data: Fabrice T. Nyambod, Murishi Onesphore, Aja Wagoh, January G. Msemakweli

Drafting of the manuscript: Fabrice T. Nyambod, Murishi Onesphore, Aja Wagoh, January G. Msemakweli

Critical review of the manuscript for important intellectual content: Fabrice T. Nyambod, Murishi Onesphore, Aja Wagoh, January G. Msemakweli

Supervision: Fabrice T. Nyambod, January G. Msemakweli

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Acknowledgements

The eICU Collaborative Research Database version 2.0 is available at https://physionet.org/content/eicu-crd/2.0/ to credentialed investigators who complete human subjects research training and sign a data use agreement that prohibits redistribution; patient-level data therefore cannot be shared by the authors. The full analysis code, including the extraction queries, and the unit-level aggregate results underlying Figures 2 and 3 are archived at https://doi.org/10.5281/zenodo.21856252 and maintained at https://github.com/fabrice-tiku-nyambod/Treatment-limitation for continued access. The authors thank the Massachusetts Institute of Technology Laboratory for Computational Physiology and Philips Healthcare for developing and maintaining the eICU Collaborative Research Database, and PhysioNet for hosting it.

References

1. Mark NM, Rayner SG, Lee NJ, Curtis JR: Global variability in withholding and withdrawal of life-sustaining treatment in the intensive care unit: a systematic review. Intensive Care Med. 2015, 41:1572-85. 10.1007/s00134-015-3810-5
2. Avidan A, Sprung CL, Schefold JC, et al.: Variations in end-of-life practices in intensive care units worldwide (Ethicus-2): a prospective observational study. Lancet Respir Med. 2021, 9:1101-10. 10.1016/S2213-2600(21)00261-7
3. Barnato AE, Tate JA, Rodriguez KL, Zickmund SL, Arnold RM: Norms of decision making in the ICU: a case

study of two academic medical centers at the extremes of end-of-life treatment intensity. Intensive Care Med. 2012, 38:1886-96. 10.1007/s00134-012-2661-6

4. Hart JL, Harhay MO, Gabler NB, Ratcliffe SJ, Quill CM, Halpern SD: Variability among US intensive care units in managing the care of patients admitted with preexisting limits on life-sustaining therapies. JAMA Intern Med. 2015, 175:1019-26. 10.1001/jamainternmed.2015.0372
5. Lee SI, Ju YR, Kang DH, Lee JE: Characteristics and outcomes of patients with do-not-resuscitate and physician orders for life-sustaining treatment in a medical intensive care unit: a retrospective cohort study. BMC Palliat Care. 2024, 23:42. 10.1186/s12904-024-01375-w
6. Ouyang DJ, Lief L, Russell D, et al.: Timing is everything: early do-not-resuscitate orders in the intensive care unit and patient outcomes. PLoS One. 2020, 15:e0227971. 10.1371/journal.pone.0227971
7. Driggers KE, Keenan LM, Alcover KC, et al.: Unintended consequences of code status in the intensive care unit: what happens after a do-not-resuscitate order is placed? A retrospective cohort study. J Palliat Med. 2024, 27:508-14. 10.1089/jpm.2023.0289
8. Quill TE, Holloway R: Time-limited trials near the end of life . JAMA. 2011, 306:1483-4. 10.1001/jama.2011.1413
9. Knaus WA, Draper EA, Wagner DP, Zimmerman JE: APACHE II: a severity of disease classification system . Crit Care Med. 1985, 13:818-29. 10.1097/00003246-198510000-00009
10. Zimmerman JE, Kramer AA, McNair DS, Malila FM: Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. Crit Care Med. 2006, 34:1297-310. 10.1097/01.CCM.0000215112.84523.F0
11. Higgins TL, Teres D, Copes WS, Nathanson BH, Stark M, Kramer AA: Assessing contemporary intensive care unit outcome: an updated Mortality Probability Admission Model (MPM0-III). Crit Care Med. 2007, 35:827-35. 10.1097/01.CCM.0000257337.63529.9F
12. Soares M, Dongelmans DA: Why should we not use APACHE II for performance measurement and benchmarking?. Rev Bras Ter Intensiva. 2017, 29:268-70. 10.5935/0103-507X.20170043
13. Sarkar R, Martin C, Mattie H, Gichoya JW, Stone DJ, Celi LA: Performance of intensive care unit severity scoring systems across different ethnicities in the USA: a retrospective observational study. Lancet Digit Health. 2021, 3:e241-9. 10.1016/S2589-7500(21)00022-4
14. Glance LG, Osler TM, Dick A: Rating the quality of intensive care units: is it a function of the intensive care unit scoring system?. Crit Care Med. 2002, 30:1976-82. 10.1097/00003246-200209000-00005
15. Kramer AA, Higgins TL, Zimmerman JE: Comparing observed and predicted mortality among ICUs using different prognostic systems: why do performance assessments differ?. Crit Care Med. 2015, 43:261-9. 10.1097/CCM.0000000000000694
16. Liang Y, Jiang X, Zhang B: Evaluating risk-adjusted hospital performance using large-scale data on mortality rates of patients in intensive care units: a flexible semi-nonparametric modeling approach. IEEE J Transl Eng Health Med. 2023, 11:232-46. 10.1109/JTEHM.2023.3257179
17. Walkey AJ, Weinberg J, Wiener RS, Cooke CR, Lindenauer PK: Association of do-not-resuscitate orders and hospital mortality rate among patients with pneumonia. JAMA Intern Med. 2016, 176:97-104. 10.1001/jamainternmed.2015.6324
18. Bertolini G, Boffelli S, Malacarne P, et al.: End-of-life decision-making and quality of ICU performance: an observational study in 84 Italian units. Intensive Care Med. 2010, 36:1495-504. 10.1007/s00134-010-1910-9
19. Feng S, Dubin JA: Identifying early-measured variables associated with APACHE IVa providing incorrect in-hospital mortality predictions for critical care patients. Sci Rep. 2021, 11:22203. 10.1038/s41598-021-01290-7
20. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP: The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. PLoS Med. 2007, 4:e296. 10.1371/journal.pmed.0040296
21. Collins GS, Reitsma JB, Altman DG, Moons KG: Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med. 2015, 162:55-63. 10.7326/M14-0697
22. Pollard TJ, Johnson AE, Raffa JD, Celi LA, Mark RG, Badawi O: The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Sci Data. 2018, 5:180178. 10.1038/sdata.2018.178
23. Austin PC, Steyerberg EW: The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. Stat Med. 2019, 38:4051-65. 10.1002/sim.8281
24. Gao S, Wang Y, Yang L, Wang Z, Huang W: Characteristics and clinical subtypes of cancer patients in the intensive care unit: a retrospective observational study for two large databases. Ann Transl Med. 2021, 9:13. 10.21037/atm-20-4634
25. Austin PC, Stuart EA: Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. Stat Med. 2015, 34:3661-79. 10.1002/sim.6607
26. Merlo J, Chaix B, Ohlsson H, et al.: A brief conceptual tutorial of multilevel analysis in social epidemiology: using measures of clustering in multilevel logistic regression to investigate contextual phenomena. J Epidemiol Community Health. 2006, 60:290-7. 10.1136/jech.2004.029454
27. Spiegelhalter DJ: Funnel plots for comparing institutional performance. Stat Med. 2005, 24:1185-202. 10.1002/sim.1970
28. Ma J, Chi S, Buettner B, et al.: Early palliative care consultation in the medical ICU: a cluster randomized crossover trial. Crit Care Med. 2019, 47:1707-15. 10.1097/CCM.0000000000004016
29. Nerenz DR, Austin JM, Deutscher D, et al.: Adjusting quality measures for social risk factors can promote equity in health care. Health Aff (Millwood). 2021, 40:637-44. 10.1377/hlthaff.2020.01764
30. Wade CE, del Junco DJ, Fox EE, et al.: Do-not-resuscitate orders in trauma patients may bias mortality-based effect estimates: an evaluation using the PROMMTT study. J Trauma Acute Care Surg. 2013, 75:S89-96. 10.1097/TA.0b013e31828fa422