Published in final edited form as:

Crit Care Med. 2022 July 01; 50(7): 1040–1050. doi:10.1097/CCM.0000000000005518.

The Global Open Source Severity of Illness Score (GOSSIS)

Jesse D. Raffa, PhD[1],

Alistair E. W. Johnson, DPhil[1],

Zach O'Brien, MBBS[2],

Tom J. Pollard, PhD[1],

Roger G. Mark, MD, PhD[1,3],

Leo A. Celi, MD, MPH, MSc[1,3],

David Pilcher, FRACP, FCICM[4,5,6],

Omar Badawi, PharmD, MPH[7]

[1]Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA.

[2]Austin Health, Melbourne, VIC, Australia.

[3]Beth Israel Deaconess Medical Center, Boston, MA.

[4]Department of Intensive Care and Hyperbaric Medicine, Alfred Hospital, Melbourne, VIC, Australia.

[5]Australian and New Zealand Intensive Care Research Centre, School of Public Health and Preventive Medicine, Monash University, Alfred Hospital, Melbourne, VIC, Australia.

[6]Centre for Outcome and Resource Evaluation, Australian and New Zealand Intensive Care Society, Melbourne, VIC, Australia.

[7]Connected Care Informatics, Philips Healthcare, Baltimore, MD.

Abstract

OBJECTIVES: To develop and demonstrate the feasibility of a Global Open Source Severity of Illness Score (GOSSIS)-1 for critical care patients, which generalizes across healthcare systems and countries.

DESIGN: A merger of several critical care multicenter cohorts derived from registry and electronic health record data. Data were split into training (70%) and test (30%) sets, using each set exclusively for development and evaluation, respectively. Missing data were imputed when not available.

For information regarding this article, jraffa@mit.edu.
New affiliation for Dr. Johnson: Peter Gilgan Center for Research and Learning, The Hospital for Sick Children, Toronto, ON, Canada.
New affiliation for Dr. Badawi: Medical Device Innovation Consortium, Arlington, VA.

Drs. Raffa's, Pollard's, and Mark's institutions received funding from Philips Healthcare. Drs. Raffa, Pollard, Mark, and Celi received support for article research from the National Institutes of Health (NIH). Dr. Johnson's institution received funding from the NIH. Dr. Badawi received funding from Philips Healthcare. The remaining authors have disclosed that they do not have any potential conflicts of interest.

SETTING/PATIENTS: Two large multicenter datasets from Australia and New Zealand (Australian and New Zealand Intensive Care Society Adult Patient Database [ANZICS-APD]) and the United States (eICU Collaborative Research Database [eICU-CRD]) representing 249,229 and 131,051 patients, respectively. ANZICS-APD and eICU-CRD contributed data from 162 and 204 hospitals, respectively. The cohort included all ICU admissions discharged in 2014–2015, excluding patients less than 16 years old, admissions less than 6 hours, and those with a previous ICU stay.

INTERVENTIONS: Not applicable.

MEASUREMENTS AND MAIN RESULTS: GOSSIS-1 uses data collected during the ICU stay's first 24 hours, including extrema values for vital signs and laboratory results, admission diagnosis, the Glasgow Coma Scale, chronic comorbidities, and admission/demographic variables. The datasets showed significant variation in admission-related variables, case-mix, and average physiologic state. Despite this heterogeneity, test set discrimination of GOSSIS-1 was high (area under the receiver operator characteristic curve [AUROC], 0.918; 95% CI, 0.915–0.921) and calibration was excellent (standardized mortality ratio [SMR], 0.986; 95% CI, 0.966–1.005; Brier score, 0.050). Performance was held within ANZICS-APD (AUROC, 0.925; SMR, 0.982; Brier score, 0.047) and eICU-CRD (AUROC, 0.904; SMR, 0.992; Brier score, 0.055). Compared with GOSSIS-1, Acute Physiology and Chronic Health Evaluation (APACHE)-IIIj (ANZICS-APD) and APACHE-IVa (eICU-CRD), had worse discrimination with AUROCs of 0.904 and 0.869, and poorer calibration with SMRs of 0.594 and 0.770, and Brier scores of 0.059 and 0.063, respectively.

CONCLUSIONS: GOSSIS-1 is a modern, free, open-source inhospital mortality prediction algorithm for critical care patients, achieving excellent discrimination and calibration across three countries.

Keywords

Acute Physiology and Chronic Health Evaluation; benchmarking; global health; mortality prediction; prediction algorithm; severity of illness

Intensive care medicine is characterized by the management of patients at the highest risk of deterioration and death (1), yet includes a heterogeneous population with substantial variation in expected outcomes (2). To account for this heterogeneity in risk, severity of illness (SOI) scores are integral to quality of care evaluations, resource management, and the stratification of patients in research (4). Several scoring systems exist but may be limited by poor generalizability beyond their derivation cohorts and are known to decrease in accuracy over time (5-7). The need remains for a high-quality SOI score that is internationally valid and thereby facilitates intensive care benchmarking on a global scale. For this reason, a consortium of investigators was formed to develop a collection of open-source ICU SOI scores; the Global Open Source Severity of Illness Score (GOSSIS). This work demonstrates the feasibility of building multinational SOI scores by leveraging existing critical care databases.

The Australian and New Zealand Intensive Care Society Adult Patient Database (ANZICS-APD) is one of the largest intensive care datasets in the world, containing high-quality patient-level data from more than 90% of ICUs in Australia and New Zealand (ANZ) (8, 9). Similarly, the eICU Collaborative Research Database (eICU-CRD) contains granular data on more than 200,000 admissions across 335 ICUs in the United States (10). We sought to harmonize these datasets so that robust international SOI scores could be developed as a proof-of-concept, with greater accuracy and external validity than scoring systems currently available. Furthermore, we aimed to outline a methodology for the future generation of global predictive models, to be used as national intensive care registries become increasingly widespread (11).

METHODS

The process of developing GOSSIS-1 is illustrated in Supplementary Figure 1 (http://links.lww.com/CCM/H89), which includes four general phases (1. data extraction, 2. data splitting, 3. model tuning, and 4. model evaluation) used to build and evaluate the GOSSIS-1 prediction algorithm for inhospital mortality.

Data Extraction

Data Sources.—Data were collected from ANZ through the ANZICS-APD (8) and the United States via the eICU-CRD v2.0 (10). Both datasets are large multicenter databases of critical care admissions with excellent representation of the heterogeneous care settings and populations throughout ANZ and the United States.

ANZICS-APD is collected quarterly as a registry of critical care admissions in the two countries, as a means to provide benchmarking of hospital performance for the contributing centers. Data collected as part of the ANZICS-APD minimal dataset include physiologic and disease characteristics of the patient over the first 24 hours of the ICU admission, along with patient and admission specific variables such as, age and ICU admission source, that are routinely collected for Acute Physiology and Chronic Health Evaluation (APACHE)-III (12) and Australian and New Zealand Risk of Death (ANZROD) (13) scoring. Additionally, the ANZICS-APD dataset collects other physiologic variables (e.g., common laboratory tests), and other demographic or patient variables. Full details, including the data collection form (14) and data dictionary (15) can be found on their website.

eICU-CRD is a critical care database collected through interfaces to laboratories, vital sign monitors, pharmacy records, and other data input into electronic health records. While these data are collected as part of routine care, protocolized collection of APACHE-IVa (16) variables are obtained for benchmarking purposes. Derivation of the eICU-CRD dataset can be found at https://github.com/MIT-LCP/GOSSIS/.

Patients were included if their ICU admission was discharged in 2014 or 2015 and excluded if they were less than 16 years old, did not have a recorded heart rate, were missing their inhospital mortality outcome, had an ICU stay less than 6 hours, or were identified with a previous ICU admission. The full patient disposition is shown in Figure 1.

This study was exempt from institutional review board approval due to the retrospective design, lack of direct patient intervention, and the security schema for which the reidentification risk was certified as meeting safe harbor standards by Privacert (Cambridge, MA) (HIPAA Certification no. 1031219-2).

Variables.—The complete set of variables included in the GOSSIS-1 dataset are described in full in Supplementary Tables 1 and 2 (http://links.lww.com/CCM/H89).

For each of the variables used in computing the Acute Physiology Score (APS) (12), the extrema (minimum and maximum) observed over the same APS observation period were collected, with one extremum generally matching the value used for APS calculation. Twenty-four-hour extrema laboratory and physiologic variables for serum calcium, hemoglobin, international normalized ratio (INR) of prothrombin time, lactate, potassium, platelets, diastolic and systolic blood pressure, and oxygen saturation were also included along with demographic and admission level variables include basic information about the hospital and ICU admission.

Data Harmonization.—The two datasets were harmonized at the Laboratory for Computational Physiology at the Massachusetts Institute of Technology. The process of harmonization involved transforming a common set of variables to correct for varying units of measurement, differing levels of recorded categories, reconciling permissible ranges by truncating the values to the most restrictive range, and combining the data into a common dataset (https://github.com/MIT-LCP/GOSSIS/).

ANZICS-APD and eICU-CRD data include the concept of admission diagnosis for APACHE-IIIj and APACHE-IVa predictions, respectively. While ANZICS-APD includes admission diagnoses, a mapping was needed to translate to the common set of diagnoses used for GOSSIS-1, resulting in 529 admission diagnoses, 117 diagnosis groupings, and 12 body system groups. A full accounting of the diagnosis groupings can be found in Supplementary Tables 12 and 13 (http://links.lww.com/CCM/H89) and the mapping can be found at https://github.com/MIT-LCP/GOSSIS/.

Transformations.—In addition to any transformations required for correcting for units of measurement, the Glasgow Coma Score and ventilation status were simplified (details in Supplementary Section 2.3, http://links.lww.com/CCM/H89). After imputation, physiologic variables with a minimum and maximum were transformed to reduce potential colinearity by taking the sum and differences of the extrema.

Data Splitting

To provide an objective assessment of GOSSIS-1's performance and prevent over-fitting, data were split randomly assigning patients to a training or a test set, representing 70% and 30% of the data, respectively. The test set was set aside until a final model had been chosen during the model tuning process. After allocating the training and test sets, the training data was again randomly split into five approximately equally sized partitions to conduct five-fold cross-validation (CV). A comparison of training and test set characteristics are available in Supplementary Table 5 (http://links.lww.com/CCM/H89).

Imputation of Missing or Incomplete Data

Missing or incomplete data occurred in both ANZICS-APD and eICU-CRD. The patterns and proportion of data missing for each variable are presented in Supplementary Figures 2-5 (http://links.lww.com/CCM/H89). For physiologic measurements, APS assumes when data were missing, the missing variables should be considered a normal value, contributing zero points. Conversely, we devised an approach that imputed missing values using a prediction model for each of GOSSIS-1's inputs as health systems and standards of care may vary across countries and generate different missingness prevalence and patterns.

Three imputation algorithms were evaluated as part of the tuning process of building GOSSIS-1 (details provided in Supplementary Section 2.5, http://links.lww.com/CCM/H89). The imputation algorithms differ in many respects, but none used any outcome data (e.g., mortality or length of stay).

Model Tuning

Model tuning was performed using five-fold CV The complete tuning process and results are presented in the Supplementary Section 2.4 (http://links.lww.com/CCM/H89). Once the candidate models were built, tuning parameters were chosen by computing measures of discrimination: area under the receiver operator characteristic curve (AUROC) (17), and measures of calibration: the Brier score (18), and standardized mortality ratio (SMR) (19). The final model was chosen by assessing the overall and dataset-specific performance across all performance metrics.

Model Fitting

Leveraging the complete dataset available after imputation, logistic regression models were fit using generalized additive mixed models by the mgcv (20) package in R. The mgcv package allows the modeling of complex relationships between the predictor variables and the outcomes using smooth functions of the numeric predictors. This approach allows for a high degree of model flexibility while maintaining the ability to interpret and explain the relationships between the predictors and the outcome.

Given that most physiologic numeric variables were available in pairs (e.g., the minimum and maximum value, parameterized as the sum and difference), we used tensor smooth products with cubic regression splines to capture the relationship between these types of variables and the outcome. Admission diagnoses were modeled through nested random effects. The dimension of the basis of the smooth terms was varied globally across all numeric variables, and performance was assessed at values 5, 8, 10, 12, 15, 17, and 20 dimensions.

Evaluation

After model tuning, the final model was trained using the full training data. The test set is processed through the same pipeline: transformation, imputation, and prediction generation. Test set performance is evaluated using the same performance metrics that guided tuning along with the Hosmer-Lemeshow goodness-of-fit test statistic (21), using 10 quantile bins. Statistically significant differences in AUROC between GOSSIS-1 and other

SOI scores were assessed using Delong's method (22) and the significance level was set at 0.05. Calibration plots were generated to assess performance within each risk decile. Evaluating the utility of GOSSIS-1 as a benchmarking tool was done by computing the hospital-specific SMR in test set patients using a funnel plot. Sensitivity analyses were conducted, focusing on the performance of the final version of GOSSIS-1 in subgroups of patients (Supplementary Tables 14-18, http://links.lww.com/CCM/H89).

RESULTS

Of the 500,881 ICU admissions obtained from 2014 to 2015, 380,280 (75.9%) were eligible for GOSSIS-1, with the majority of those admissions excluded being due to having a short duration (< 6 hr) or an ineligible nonindex ICU stay (e.g., readmission to the ICU). The admission characteristics of ANZICS-APD and eICU-CRD are compared in Table 1, where clear differences between the two datasets are observed. Notably, while specialty ICUs exist in the United States, ANZ have no such units. Further comparison of clinical characteristics can be found in Supplementary Tables 6 and 7 (http://links.lww.com/CCM/H89). Although most candidate models had good discrimination and calibration in the validation samples, a final model of imputation algorithm 3, tuning length 2 and k = 20 was selected to be the final model as it was among the top three best-performing candidate models across all metrics and datasets (see Supplementary Fig. 7 and Supplementary Section 3.2, http://links.lww.com/CCM/H89, for details).

Evaluation

A description of the final model can be found in the Supplementary Section 3.3 (http://links.lww.com/CCM/H89) with the performance in the test set presented in Table 2. In all performance metrics and datasets, GOSSIS-1 was found to be superior to APACHE-IIIj and APACHE-IVa. A comparison of the AUROC GOSSIS-1 to the APACHE alternative (IIIj in ANZICS-APD and IVa for eICU-CRD) was found to be statistically significant for comparisons within each cohort and overall ($p < 0.001$ for all pairwise comparisons).

A calibration plot of GOSSIS-1 and APACHE-IIIj or APACHE-IVa is presented in Figure 2. The decile level calibration is superior to APACHE's in all deciles of risk, particularly among patients with the highest predicted risk of inhospital death.

A funnel plot of the hospital-level SMR versus the hospital sample size is presented in Figure 3. Overall, 17 of 365 hospitals (4.7%) have SMRs that exceed the upper or lower bounds of the 99% CI, with 13 of 161 (8.1%) and four of 204 (2%) meeting that criteria in ANZICS-APD and eICU-CRD, respectively. Further evaluation of GOSSIS-1 at the hospital- and subgroup-level are discussed in Supplementary Section 3.4 (http://links.lww.com/CCM/H89).

DISCUSSION

Key Findings

We have demonstrated that the harmonization of heterogeneous, international intensive care databases is feasible. We report a robust methodology with which to extract and split data from such a database and subsequently tune and evaluate a predictive model. With this methodology, we have developed GOSSIS-1, a SOI score for the prediction of inhospital mortality among patients admitted to the ICU. Our results indicate that GOSSIS-1 offers superior discrimination and calibration in our cohort when compared with the commonly used scoring systems, performed consistently well across patient subgroups (see Supplementary Table 14, http://links.lww.com/CCM/H89) and may provide a basis for benchmarking across countries.

Relationship With Previous Studies

Outcome prediction scores typically estimate the expected inhospital mortality and allow for the calculation of SMRs, so ICUs may be compared for benchmarking purposes and outliers identified (23, 24). The most widely used SOI scoring systems are the APACHE, Mortality Prediction Model (MPM), and the Simplified Acute Physiology Score (SAPS), each with several iterations (16, 25, 26). The APACHE scoring systems were developed using patients entirely from the United States and have consistently been reported to demonstrate better discrimination than other systems (7, 16, 27-29). However, several studies have shown that the latest APACHE-IV score has poor calibration when applied outside of the United States (7, 30, 31). Indeed, we note in this study (Fig. 2B), the lack of calibration of APACHE scores is much more severe in ANZICS-APD than in eICU-CRD. Similarly, the MPM III score was derived from more than 120,000 patients in the United States, although again has poor calibration when applied to cohorts in other countries (7). In contrast, SAPS 3 was derived using data on patients from five continents and also provides several equations allowing it to be customized to particular geographical regions of the world (26, 32). While this made it a promising scoring system for international benchmarking, it has been reported to have lower discrimination and calibration than other scores when applied to various countries (33, 34) perhaps because it was derived from a comparatively small sample of fewer than 20,000 patients, with non-European countries poorly represented (3, 29). SAPS 3 also differs in that it collects fewer variables over a shorter data collection window (first hour).

GOSSIS-1 has many similarities to APACHE III and IV, and current users of these scoring systems may find that they already are collecting many of the necessary data elements to use GOSSIS-1. Additionally, GOSSIS-1 uses similar inclusion criteria and data collection windows (a patient’s first 24 hr of ICU admission) to APACHE. On the other hand, there are many substantive differences. GOSSIS-1 adds a new set of physiologic variables (e.g., serum calcium, hemoglobin, INR, lactate, potassium, and platelets), while removing others (e.g., urine output) and uses both the minimum and maximum physiologic variables to train the model, not only the “worst” physiologic state but also the range of values these variables have been observed at. The modeling approach differs in many respects as well, as the two sets of models take a different approach to the handling of missing data, GOSSIS-1 does not use the APS point system, and while APACHE has separate models for coronary artery bypass graft (CABG) and non-CABG patients, GOSSIS-1 does not.

Implications of Findings

Our study demonstrates that intensive care databases from multiple countries can be harmonized despite significant differences in the data structure, patients, and the healthcare systems that they represent. Our findings imply that machine/statistical learning methods are effective for imputing missing data in a harmonized database and that our methods for developing and evaluating a predictive model are advantageous. The results of our model evaluation imply that GOSSIS-1 may demonstrate improved discrimination and calibration when compared with SOI scores in current use and that this may be maintained when applied to patients in different countries. Consequently, GOSSIS-1 may allow for the standardized comparison of patient outcomes between countries, at both a population level and within specific patient cohorts.

Limitations

As can be found in Table 1 and Supplementary Tables 6 and 7 (http://links.lww.com/CCM/H89), the clinical and demographic characteristics of the patient populations among the two cohorts differ considerably. Additionally, we know that there are substantial differences in how the units dedicated to critical care are structured in the three countries, especially considering the U.S. cohort represents tele-critical systems. When compared with the United States, ANZ critical care units generally have higher nursing-to-patient ratios, lack dedicated respiratory therapists, and are more likely to be closed units that may affect patient outcomes (35). The consistent performance of GOSSIS-1 in both ANZ and U.S. hospitals lends credence that the approach may be successful when trained on datasets that may not only have a heterogeneous case-mix of patients but also differences in the overall structure of critical care or healthcare as a whole in other regions as well. Our study also has potential limitations. As with all predictive models, if it is to be used in a population outside that used for its derivation, it should be independently validated in that context first. While GOSSIS-1 may have reduced accuracy and may not exhibit better performance compared with alternatives when applied more broadly, we suspect that the large, heterogeneous population used to develop the score will improve its generalizability. Low- and middle-income countries were not represented in the databases that were available to us, and the validity of GOSSIS-1 in these countries is unknown though future iterations are expected to expand representation. At the time of publication, we were not able to externally validate GOSSIS-1 in an independent cohort of critical care patients though external validation is underway in several diverse cohorts. The limitations of GOSSIS-1 identified in these studies will inform the direction of the consortium's efforts to develop future SOI scores. Such studies may identify other limitations including overfitting and sensitivity to case-mix. Additionally, it should be acknowledged that APACHE-IV has updated weights that may improve its performance in this cohort, but this work is not publicly available, and we cannot verify this.

Additionally, we only included data from patients' first ICU admission and when the ICU stay was greater than or equal to 6 hours. Therefore, predictions on patients with prolonged critical illness requiring multiple ICU admissions and those with extremely high or low SOI may have reduced validity. Some SOI models choose to exclude or build distinct models for certain subsets of patients (e.g., CABG patients in APACHE-IV), while GOSSIS-1 chose a single model for all eligible patients. GOSSIS-1 could perform worse than other SOI scores in these subpopulations, but this was generally not observed when GOSSIS-1 was evaluated on test set patients from these subsets. Further, GOSSIS-1 has been developed using data from patients treated in 2014–2015. As such, the model may have reduced performance when applied to patients today and the validity of GOSSIS-1 during the COVID-19 pandemic is unknown.

The methodology outlined in this article describes a complete pipeline we used when creating GOSSIS-1. Alternative approaches to modeling these data certainly exist, and our choices were selected to emphasize interpretability and explainability of the model, while addressing the challenges we encountered and believe we will encounter when developing future versions of GOSSIS. A substantial challenge in the development of GOSSIS-1 was the significant amount of missing data. We developed a robust and objective method for imputing missing variables and did not observe any decreases in internal validity due to missing data. The imputation approach may also be useful for reconciling differences in the variables collected by different cohorts we used when building future versions of GOSSIS. Our imputation approach may help avoid unpalatable alternatives, such as not using an entire dataset because it is missing a variable, or dropping the missing variable from all datasets. Our modeling of diagnosis using nested random effects may allow for easier integration of diagnosis data by pooling strength from similar diagnoses and allowing rare and new diagnoses to be added without much difficulty.

CONCLUSIONS

We successfully harmonized two large intensive care databases from three countries. In doing so, we were able to develop GOSSIS-1, an open-source SOI score that may offer superior discrimination and calibration than SOI scores in common use. Due to the heterogeneity of patients used to develop GOSSIS-1, it may offer greater generalizability than other systems, although this requires future validation. Furthermore, we report a high-quality and reproducible methodology for the harmonization of international medical databases and the development of MPMs.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

REFERENCES

1. Marshall JC, Bosco L, Adhikari NK, et al. : What is an intensive care unit? A report of the task force of the World Federation of Societies of Intensive and Critical Care Medicine. J Crit Care 2017; 37:270–276 [PubMed: 27612678]

2. Adhikari NK, Fowler RA, Bhagwanjee S, et al. : Critical care and the global burden of critical illness in adults. Lancet 2010; 376:1339–1346 [PubMed: 20934212]

3. Keegan MT, Gajic O, Afessa B: Severity of illness scoring systems in the intensive care unit. Crit Care Med 2011; 39:163–169 [PubMed: 20838329]

4. Vincent JL, Moreno R: Clinical review: Scoring systems in the critically ill. Crit Care 2010; 14:207 [PubMed: 20392287]

5. Paul E, Bailey M, Van Lint A, et al. : Performance of APACHE III over time in Australia and New Zealand: A retrospective cohort study. Anaesth Intensive Care 2012; 40:980–994 [PubMed: 23194207]

6. Strand K, Flaatten H: Severity scoring in the ICU: A review. Acta Anaesthesiol Scand 2008; 52:467–478 [PubMed: 18339152]

7. Nassar AP Jr, Mocelin AO, Nunes AL, et al. : Caution when using prognostic models: A prospective comparison of 3 recent prognostic models. J Crit Care 2012; 27:423.e1–e7

8. Stow PJ, Hart GK, Higlett T, et al. ; ANZICS Database Management Committee: Development and implementation of a high-quality clinical database: The Australian and New Zealand Intensive Care Society Adult Patient Database. J Crit Care 2006; 21:133–141 [PubMed: 16769456]

9. ANZICS Centre for Outcome and Resource Evaluation: 2018 ANZICS CORE Report. 2018. Available at: https://www.anzics.com.au/wp-content/uploads/2019/10/2018-ANZICS-CORE-Report.pdf. Accessed March 14, 2022

10. Pollard TJ, Johnson AEW, Raffa JD, et al. : The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Sci Data 2018; 5:180178 [PubMed: 30204154]

11. Vijayaraghavan BKT, Venkatraman R, Ramakrishnan N: Critical care registries: The next big stride? Indian J Crit Care Med 2019; 23:387 [PubMed: 31485112]

12. 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 [PubMed: 1959406]

13. Paul E, Bailey M, Kasza J, et al. : The ANZROD model: Better benchmarking of ICU outcomes and detection of outliers. Crit Care Resusc 2016; 18:25–36 [PubMed: 26947413]

14. ANZICS Centre for Outcome and Resource Evaluation: APD Data Collection Form, Version 22.0. 2022. Available at: https://www.anzics.com.au/wp-content/uploads/2021/03/APD-Data-Collection-Form.pdf. Accessed March 14, 2022

15. ANZICS Centre for Outcome and Resource Evaluation: APD Data Dictionary, Version 6.0. 2022. Available at: https://www.anzics.com.au/wp-content/uploads/2021/03/ANZICS-APD-Dictionary.pdf. Accessed March 14, 2022

16. 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 [PubMed: 16540951]

17. Hanley JA, McNeil BJ: A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983; 148:839–843 [PubMed: 6878708]

18. Hunt T: ModelMetrics: Rapid Calculation of Model Metrics. 2020. Available at: https://cran.r-project.org/package=ModelMetrics. Accessed December 22, 2021

19. Breslow NE: Statistical methods in cancer research II: The design and analysis of cohort studies. IARC Scientific Publish 1987; 82:1–406

20. Wood S: Generalized Additive Models: An Introduction With R. Second Edition. Boca Raton, FL, CRC Press, 2017

21. Hosmer DW, Lemeshow S: Goodness of fit tests for the multiple logistic regression model. Commun Stat Theory Methods 1980; 9:1043–1069

22. Robin X, Turck N, Hainard A, et al. : pROC: An open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011; 12:77 [PubMed: 21414208]

23. Capuzzo M, Ranzani OT: How objective is the observed mortality following critical care? Intensive Care Med 2013; 39:2047–2049 [PubMed: 23982727]

24. Flaatten H: The present use of quality indicators in the intensive care unit. Acta Anaesthesiol Scand 2012; 56:1078–1083 [PubMed: 22339772]

25. Higgins TL, Teres D, Copes WS, et al. : Assessing contemporary intensive care unit outcome: An updated Mortality Probability Admission Model (MPM0-III). Crit Care Med 2007; 35:827–835 [PubMed: 17255863]

26. 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 [PubMed: 16132892]

27. 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 [PubMed: 18403657]

28. Kramer AA, Higgins TL, Zimmerman JE: 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 [PubMed: 24158174]

29. 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 [PubMed: 22499827]

30. Brinkman S, Bakhshi-Raiez F, Abu-Hanna A, et al. : External validation of Acute Physiology and Chronic Health Evaluation IV in Dutch intensive care units and comparison with Acute Physiology and Chronic Health Evaluation II and Simplified Acute Physiology Score II. J Crit Care 2011; 26:105. e11–e18

31. Lee H, Shon YJ, Kim H, et al. : Validation of the APACHE IV model and its comparison with the APACHE II, SAPS 3, and Korean SAPS 3 models for the prediction of hospital mortality in a Korean surgical intensive care unit. Korean J Anesthesiol 2014; 67:115–122 [PubMed: 25237448]

32. Moralez GM, Rabello LSCF, Lisboa TC, et al. ; ORCHESTRA Study Investigators: External validation of SAPS 3 and MPM0-III scores in 48,816 patients from 72 Brazilian ICUs. Ann Intensive Care 2017; 7:53 [PubMed: 28523584]

33. Poole D, Rossi C, Anghileri A, et al. : External validation of the Simplified Acute Physiology Score (SAPS) 3 in a cohort of 28,357 patients from 147 Italian intensive care units. Intensive Care Med 2009; 35:1916–1924 [PubMed: 19685038]

34. Metnitz B, Schaden E, Moreno R, et al. ; ASDI Study Group: Austrian validation and customization of the SAPS 3 Admission Score. Intensive Care Med 2009; 35:616–622 [PubMed: 18846365]

35. Sakr Y, Moreira CL, Rhodes A, et al. ; Extended Prevalence of Infection in Intensive Care Study Investigators: The impact of hospital and ICU organizational factors on outcome in critically ill patients: Results from the Extended Prevalence of Infection in Intensive Care study. Crit Care Med 2015; 43:519–526 [PubMed: 25479111]

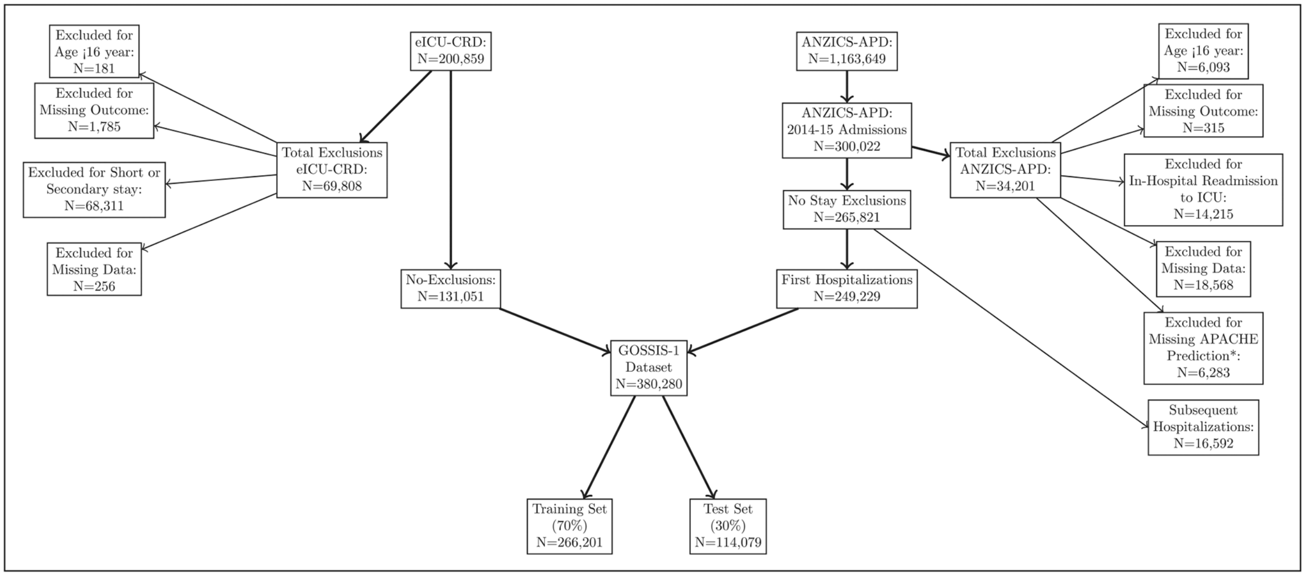

Figure 1.

Patient disposition for Global Open Source Severity of Illness Score (GOSSIS) Version 1.0 Dataset Creation. Total exclusions are broken down into reasons nonexclusively, and the same patient may be in multiple groups. *Reasons for Australian and New Zealand Intensive Care Society Adult Patient Database (ANZICS-APD) patients missing Acute Physiology and Chronic Health Evaluation (APACHE) prediction are most often due to stay a less than 6 hr, the unit not being an ICU (e.g., a high-dependency unit admission), or missing APACHE admission diagnosis. eICU-CRD = eICU Collaborative Research Database.

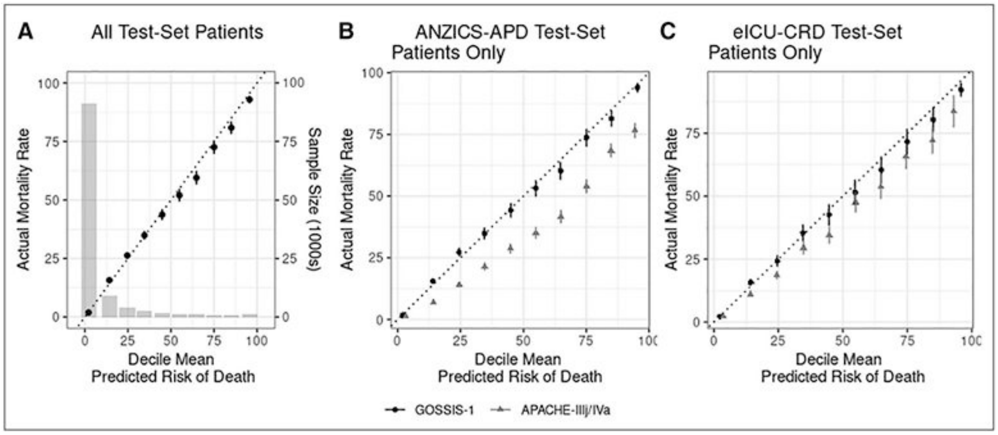

Figure 2.

Calibration plot for all test set patients (**A**), Australian and New Zealand Intensive Care Society Adult Patient Database (ANZICS-APD) test set patients (**B**), and eICU Collaborative Research Database (eICU-CRD) patients (**C**) with a valid Acute Physiology and Chronic Health Evaluation (APACHE)-IVa score. *Points* indicate the actual inhospital death rate at the decile's mean predicted risk of death for the individual scores (*black circles*: Global Open Source Severity of Illness Score [GOSSIS]-1; *gray triangles*: APACHE-IIIj [ANZICS–APD]/APACHE-IVa [eICU-CRD]). The *gray bars* on **A** represent the number of test set patients in thousands in that risk decile as predicted by GOSSIS-1.

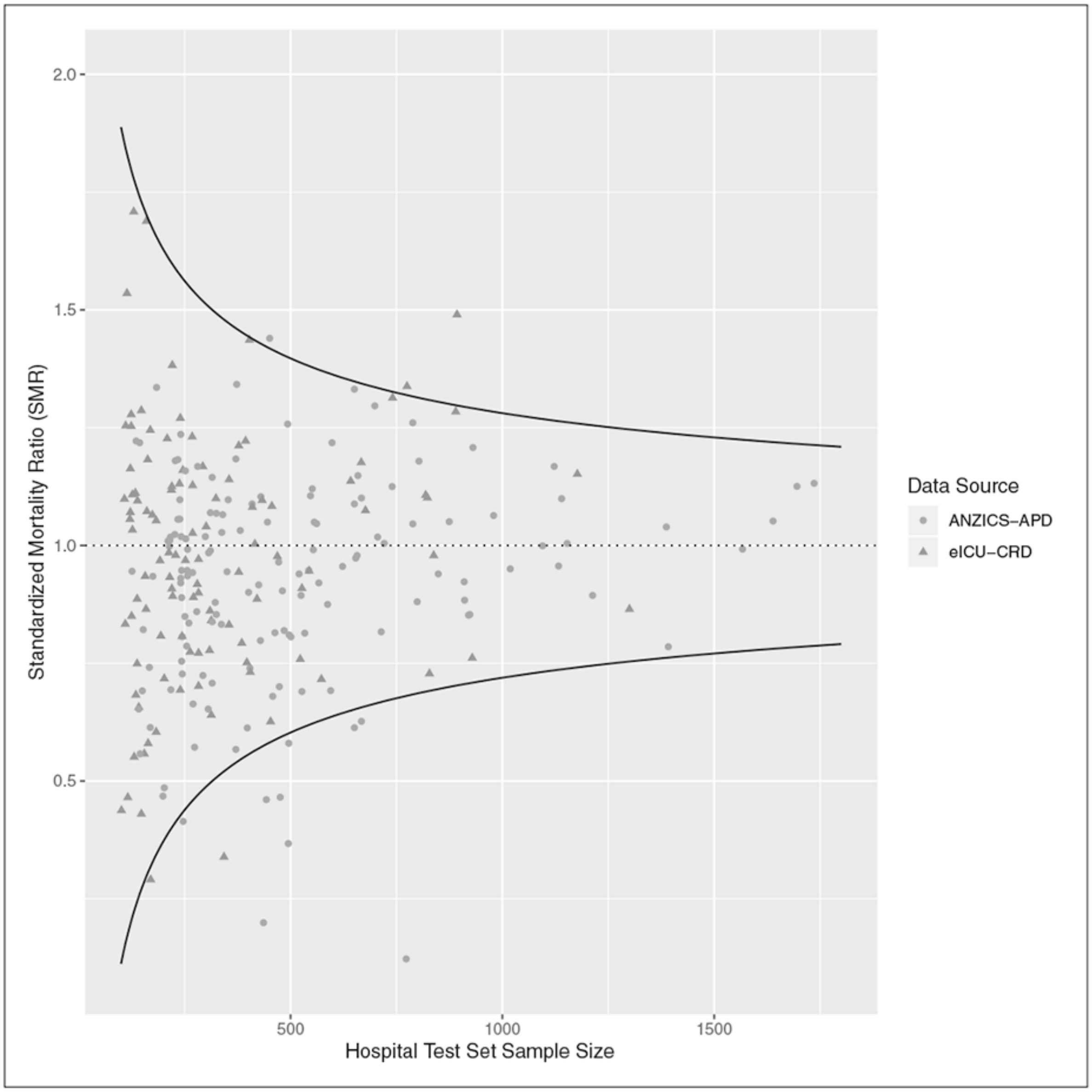

Figure 3.

Funnel plot of standardized mortality ratios (SMRs) computed using Global Open Source Severity of Illness Score-1 and test set sample size by hospital. *Shapes* indicate the data source. Hospitals contributing fewer than test set 100 hospitals admissions in 2014–2015 were excluded from the plot. The *solid lines* represent 99% piecewise confidence bands for a given sample size when the true SMR = 1.0. ANZICS-APD = Australian and New Zealand Intensive Care Society Adult Patient Database, eICU-CRD = eICU Collaborative Research Database.

TABLE 1.

Patient Characteristics Overall by Data Origin

Variable	**Group**	**Australian and New Zealand Intensive Care Society Adult Patient Database (*n* = 249,229)**	**eICU-CRD (*n* = 131,051)**	***p***
Age, median (IQR)		65.00 (50.80–75.30)	64.00 (52.00–75.00)	0.480
Gender (%)	Male	143,361 (57.5)	70,983 (54.2)	< 0.001
Country (%)	Australia	225,310 (90.4)	0 (0.0)	< 0.001
	New Zealand	23,919 (9.6)	0 (0.0)	
	United States	0 (0.0)	131,051 (100.0)	
ICU admission source (%)	Other or unknown	142 (0.1)	241 (0.2)	< 0.001
	Accident and emergency	65,341 (26.2)	77,197 (58.9)	
	Floor	32,793 (13.2)	21,583 (16.5)	
	Operating room/recovery	135,006 (54.2)	27,514 (21.0)	
	Other hospital	13,483 (5.4)	3,475 (2.7)	
	ICU to ICU transfer within the same hospital	2,464 (1.0)	1,041 (0.8)	
Type of ICU (%)	General ICU	249,229 (100.0)	0 (0.0)	< 0.001
	Cardiac ICU	0 (0.0)	8,337 (6.4)	
	Coronary care unit, cardiothoracic ICU, or cardiac surgery ICU	0 (0.0)	20,811 (15.9)	
	Medical-Surgical ICU	0 (0.0)	71,972 (54.9)	
	Medical ICU	0 (0.0)	11,010 (8.4)	
	Neuro-ICU	0 (0.0)	10,616 (8.1)	
	Surgical ICU	0 (0.0)	8,305 (6.3)	
APACHE III diagnosis body system (%)	Cardiovascular	63,545 (25.5)	41,396 (31.6)	< 0.001
	Gastrointestinal	40,757 (16.4)	12,977 (9.9)	
	Genitourinary	9,569 (3.8)	2,992 (2.3)	
	Gynecological	3,759 (1.5)	437 (0.3)	
	Hematological	1,116 (0.4)	830 (0.6)	
	Metabolic	16,855 (6.8)	10,753 (8.2)	
	Musculoskeletal/skin	14,449 (5.8)	1,669 (1.3)	
	Neurologic	31,036 (12.5)	17,885 (13.6)	

Variable	**Group**	**Australian and New Zealand Intensive Care Society Adult Patient Database (*n* = 249,229)**	**eICU-CRD (*n* = 131,051)**	***p***
	Other medical disorders	1,359 (0.5)	2,579 (2.0)	
	Respiratory	37,107 (14.9)	16,894 (12.9)	
	Sepsis	16,579 (6.7)	16,402 (12.5)	
	Trauma	13,098 (5.3)	6,237 (4.8)	
Death during hospital admission (%)		20,283 (8.1)	11,841 (9.0)	< 0.001
APACHE hospital death probability (expected deaths, mean probability)	APACHE IIIj (Australian and New Zealand Intensive Care Society Adult Patient Database) and APACHE IVa (eICU-CRD)	33,673.4 (13.5)	13,619 (11.9)	NA
Death during ICU admission (%)		13,165 (5.3)	7,317 (5.6)	< 0.001
Hospital length of stay (d), median (IQR)		8.14 (4.66–14.77)	4.86 (2.65–8.53)	< 0.001
ICU length of stay (d), median (IQR)		1.74 (0.91–3.19)	1.77 (0.96–3.22)	< 0.001
APACHE III score, mean (SD)		53.13 (25.32)	55.08 (25.64)	< 0.001
Training/test set inclusion (%)	Test	74,761 (30.0)	39,318 (30.0)	0.977

APACHE = Acute Physiology and Chronic Health Evaluation, eICU-CRD = eICU Collaborative Research Database, IQR = interquartile range.

TABLE 2.

Global Open Source Severity of Illness Score-1 Test Set Performance Versus Acute Physiology and Chronic Health Evaluation-IIIj and Acute Physiology and Chronic Health Evaluation-IVa

Prediction Algorithm (Test Cohort)	*n*	Observed Death Rate, %	Average Predicted Death Rate, %	Area Under the Receiver Operator Characteristic Curve (95% CI)	Standardized Mortality Ratio (95% CI)	Brier Score	Hosmer-Lemeshow Test Statistic (8 Degrees of Freedom)
GOSSIS-1 (overall)	114,079	8.4	8.5	0.918 (0.915–0.921)	0.986 (0.966–1.005)	0.050	103.32
GOSSIS-1 (ANZICS-APD)	74,761	8.1	8.2	0.925 (0.922–0.928)	0.982 (0.957–1.007)	0.047	87.77
GOSSIS-1 (eICU-CRD)[a]	39,318	9	9.1	0.904 (0.9–0.909)	0.992 (0.959–1.024)	0.055	25.14[b]
APACHE-IIIj (ANZICS-APD)	74,761	8.1	13.6	0.904 (0.9–0.908)	0.594 (0.579–0.609)	0.059	2,918.60
APACHE-IVa (eICU-CRD)[a]	34,398	9.1	11.8	0.869 (0.863–0.876)	0.77 (0.743–0.797)	0.063	338.71[b]

ANZICS-APD = Australian and New Zealand Intensive Care Society Adult Patient Database, APACHE = Acute Physiology and Chronic Health Evaluation, eICU-CRD = eICU Collaborative Research Database, GOSSIS-1 = Global Open Source Severity of Illness Score-1.

[a] The inclusion criteria for GOSSIS-1 and APACHE-IVa differ slightly due to the imputation approach used in GOSSIS-1. In particular, the GOSSIS-1 and the APACHE-IVa test sets differ by 4,920 patients (all among eICU-CRD patients). The excluded patients—those without an APACHE-IVa prediction were excluded because they had missing data related to non-Acute Physiology Score 3 (e.g., Glasgow Coma Score and chronic comorbidities) related components of the APACHE-IVa prediction equation. When compared on the same set of patients (i.e., those with an APACHE-IVa inhospital mortality prediction), the results remained relatively unchanged, with GOSSIS-1 achieving an area under the receiver operator characteristic curve (AUROC) of 0.906 (95% CI, 0.901–0.911), Brier score of 0.054, and standardized mortality ratio (SMR) of 0.997 (95% CI, 0.962–1.032). In the 4,920 patients without an APACHE-IVa score, there were no significant reductions in discriminative performance (AUROC, 0.893; 95% CI, 0.878–0.907) nor calibration (Brier score, 0.059; SMR, 0.957; 95% CI, 0.866–1.047) of GOSSIS-1.

[b] Calculated on a common set of 34,398 patients who had both GOSSIS-1 and APACHE IVa scores.