Predicting outcome in critical care: past, present and future

Jeremy M. Kahn[a,b]

The desire to predict the future is a near universal human trait. For centuries humans have tried to forecast future events using both natural [1] and supernatural [2] means. The field of critical care medicine is no exception. Indeed, for several reasons the ICU has led the way in clinical outcome prediction among the medical disciplines. Compared with other areas of medicine, the ICU is a data-rich environment, with data on past patients readily available for use to predict outcomes for future patients. Additionally, in the ICU, life and death decisions are made on a daily basis. Accurate outcome prediction is extremely useful for improving decision-making under uncertainty, particularly when the stakes are so high. Because of these factors, it is no surprise that ICU clinicians helped pioneer the use of outcomes predictions, applying the tools of clinical epidemiology to robustly predict outcomes of critically ill patients [3].

Yet, despite all the work in this area, ICU outcome prediction has in many ways failed to achieve its full promise. Efforts to incorporate real-time prediction to improve decision-making have not been successful [4], and audits that compare observed outcomes with predicted outcomes are generally ineffective for quality improvement [5]. This is not to say that work in this area was without merit, only that it is incomplete. Outcome prediction remains a powerful tool to inform clinical decision-making and accelerate quality improvement. We just need new and innovative ways to apply these tools, taking advantage of advances in data collection, statistical modeling, and outcomes measurement in critical care [6].

In this section of *Current Opinion in Critical Care,* we are fortunate to have international experts in the field of critical care outcomes prediction provide a roadmap for this process. The authors tackle the past, present, and future of ICU outcomes prediction, discussing where we came from, where we are now, and where we are going. Each author brings a unique perspective. Their areas of expertise are broad and include clinical epidemiology, biostatistics, sociology, and information technology. Yet, all are also directly involved with the day-to-day practice of critical care medicine, bringing real-world, first-hand experience to the discussion.

In the first article, Drs Sarah Power and David Harrison, statisticians at the United Kingdom Intensive Care National Audit and Research Center, provide the rationale for why it is valuable to predict ICU outcomes in the first place. The most important reason, of course, is to inform clinical decision-making and quality improvement through benchmarking – the systematic measurement of outcomes – which depends on accurate outcome prediction to ensure apples to apples comparisons. However, Power and Harrison also discuss some innovative emerging applications for risk prediction, including increasing the precision of randomized controlled trials and informing observational outcomes studies.

In the second article, Drs Jack Zimmerman and Andrew Kramer provide a first-hand account of the history of outcome prediction in the ICU. Dr Zimmerman was literally present at the birth of modern ICU outcome prediction. Along with Bill Knaus, Doug Wagner, and Betty Draper, Dr Zimmerman pioneered the use of physiology on presentation to predict patient outcomes in the 1970s and 1980s, positioning the ICU at the forefront of risk-measurement in health medicine through the development of the Acute Physiology and Chronic Health Evaluation (APACHE) system. Dr Kramer was the lead biostatistician responsible for stewarding APACHE into the modern era. Their review discusses the key decisions made and lessons learned along the way, and in doing so provides insight into the future of ICU outcome prediction, including use of

[a]CRISMA Center, Department of Critical Care Medicine, University of Pittsburgh School of Medicine and [b]Department of Health Policy and Management, University of Pittsburgh Graduate School of Public Health, Pittsburgh, Pennsylvania, USA

Correspondence to Jeremy M. Kahn, MD MS, Associate Professor of Critical Care Medicine and Health Policy and Management, University of Pittsburgh, Scaife Hall Room 602-B, 3550 Terrace Street, Pittsburgh, PA 15263, USA. Tel: +1 412 683 7601; e-mail: kahnjm@upmc.edu

Curr Opin Crit Care 2014, 20:542–543

DOI:10.1097/MCC.0000000000000140

physiology-based models for national benchmarking and the application of big data to critical care. As the old adage goes, 'you can't know where you are going until you know where you've been'.

In the third article, Drs Jorge Salluh and Márcio Soares describe the current state of ICU risk prediction systems, outlining the strengths and limitations of the three most common systems: APACHE, the Simplified Acute Physiology Score, and the Mortality Prediction Model (MPM). Drawing on their experience in a large, nationally representative sample of Brazilian hospitals, Drs Salluh and Soares highlight the need to continually update and recalibrate prediction models, and when possible customize them to local populations. Their review provides ICU clinicians with a practical roadmap for implementing outcome prediction in their clinical practice.

In the fourth article, Dr Christopher Cox *et al.* address one of the major limitations of the existing ICU risk prediction systems: they lack patient-centeredness. APACHE, the Simplified Acute Physiology Score, and the Mortality Prediction Model predict short-term mortality after critical illness. However, patients actually care about so much more than short-term mortality – they want information on long-term mortality and functional outcomes. Thus, the next generation of ICU outcome prediction should answer not only the question 'will this patient be alive at hospital discharge?', but also the questions 'will this patient be alive 1 year from now, and if so, what will be their quality of life?'. In this review, the authors provide a roadmap for the next generation of ICU outcome prediction scores, which will use self-reported outcomes and integrated electronic health records to provide this important information.

In the fifth review, Dr Leo Celi from the Massachusetts Institute of Technology discusses another major limitation of existing ICU risk-prediction systems: they do not fully take advantage of the multitudes of data available in the modern ICU. Our capacity to collect, store, manage, and analyze health data has evolved dramatically in the last decade. Yet, we have not fully realized the vision of using these data for risk-prediction at the bedside. Dr Celi *et al.* outline the complex issues involving the smarter use of data, providing a framework for an optimal data system that can meaningfully inform decision-making in real time. This in part is the vision of the United States Institute of Medicine's 'learning healthcare system', in which patient data are used intelligently to inform evidence-based yet personalized clinical decisions [7].

Together, these reviews provide a holistic overview of ICU outcomes prediction, taking the lessons learned from past efforts and putting them in the context of future work. Hopefully, these reviews will provide practicing clinicians with the knowledge and insight necessary for intelligent application of ICU outcome prediction tools. At the same time, these reviews should inspire future clinicians and scientists to develop new and better outcome prediction tools. Outcome prediction will always be a cornerstone of the practice of critical care medicine. Our task is to build on its strengths, understand its limitations, and not be satisfied with status quo.

Acknowledgements

Dr Kahn has received consulting fees from the United States Department of Veterans Affairs for consulting on the topic of ICU telemedicine. His institution receives grant funding from the National Institutes of Health, the United States Health Resources and Services Administration, and the Gordon and Betty Moore Foundation; and receives in-kind research support from the Cerner Corporation.

Conflicts of interest

There are no conflicts of interest.

REFERENCES

1. Richardson LF. Weather prediction by numerical process. Cambridge: Cambridge University Press; 1922.
2. Leoni E. Nostradamus and his prophecies. Mineola, NY: PN Dover; 2000.
3. Shoemaker WC, Pierchala C, Chang P, State D. Prediction of outcome and severity of illness by analysis of the frequency distributions of cardiorespiratory variables. Crit Care Med 1977; 5:82–88.
4. A controlled trial to improve care for seriously ill hospitalized patients. The study to understand prognoses and preferences for outcomes and risks of treatments (SUPPORT). The SUPPORT Principal Investigators. JAMA 1995; 274:1591–1598.
5. Van der Veer SN, de Vos MLG, van der Voort PH, *et al.* Effect of a multifaceted performance feedback strategy on length of stay compared with benchmark reports alone: a cluster randomized trial in intensive care*. Crit Care Med 2013; 41:1893–1904.
6. Kahn JM, Fuchs BD. Identifying and implementing quality improvement measures in the intensive care unit. Curr Opin Crit Care 2007; 13:709–713.
7. Olsen L, Aisner D, McGinnis JM. Institute of Medicine (US) Roundtable on Evidence-Based Medicine: The Learning Healthcare System: Workshop Summary. Washington, DC: National Academies Press (US); 2007.

A history of outcome prediction in the ICU

Jack E. Zimmerman[a,b] *and Andrew A. Kramer*[b]

Purpose of review
There are few first-hand accounts that describe the history of outcome prediction in critical care. This review summarizes the authors' personal perspectives about the development and evolution of Acute Physiology and Chronic Health Evaluation over the past 35 years.

Recent findings
We emphasize what we have learned in the past and more recently our perspectives about the current status of outcome prediction, and speculate about the future of outcome prediction.

Summary
There is increasing evidence that superior accuracy in outcome prediction requires complex modeling with detailed adjustment for diagnosis and physiologic abnormalities. Thus, an automated electronic system is recommended for gathering data and generating predictions. Support, either public or private, is required to assist users and to update and improve models. Current outcome prediction models have increasingly focused on benchmarks for resource use, a trend that seems likely to increase in the future.

Keywords
ICU, patient outcome assessment, quality assessment, risk adjustment, severity of illness index

INTRODUCTION

The standardized mortality ratio (SMR), which is the ratio of observed to predicted mortality, is the most commonly used measure of ICU quality. In Western Europe, SMR is mandated by six countries [1]. In the United States, SMRs are mandatory in all Veterans Administration ICUs [2], but are used in only 10–15% of other ICUs [3,4]. This article presents a first-hand description of the history of the Acute Physiology and Chronic Health Evaluation (APACHE) system. By focusing on our experiences and the lessons learned over the past 35 years, we hope the reader will gain insight about the past, present, and future of ICU outcome prediction.

THE 1970S: THE INFANCY OF CRITICAL CARE OUTCOME PREDICTION

Many coronary care units and ICUs were established in the United States during the 1960s. Because of the inability to adjust for patient differences, physicians working in these units found it difficult to demonstrate improvements in survival. After my residency, I (J.E.Z.) served as an internist aboard a Navy hospital ship off Vietnam in 1969; my practice there focused on the 20-bed ICU. This experience led to my return to Bethesda Naval Hospital to establish a multidisciplinary ICU. In 1972, I completed my military obligation and joined the ICU staff at George Washington University.

In 1977, William 'Bill' Knaus, a former internal medicine resident, became an ICU fellow. Bill was completing a Robert Wood Johnson fellowship and believed that ICUs represented a technology consisting of people and machines and wanted to study their efficacy. Bill subsequently joined our ICU staff, obtained a grant to assess severity of illness, and established an ICU research team: Bill Knaus, Jack Zimmerman, Douglas 'Doug' Wagner, and Elizabeth 'Betty' Draper. Bill's goal was to minimize human judgment and develop an objective, mathematical measure of severity. From 1979 to 1983, our ICU became the focus of research that resulted in the basic concepts of the APACHE system [5].

[a]Anesthesiology and Critical Care Medicine, The George Washington University, Washington, District of Columbia and [b]Cerner Corporation, Vienna, Virginia, USA

Correspondence to Andrew A. Kramer, PhD, Senior Research Manager, 1953 Gallows Road, Suite 500, Vienna, VA 22182, USA. Tel: +1 703 245 8147; fax: +1 936 7447; e-mail: akramer@cerner.com

Curr Opin Crit Care 2014, 20:550–556
DOI:10.1097/MCC.0000000000000138

KEY POINTS

- Complex prognostic models predict mortality more accurately than simplified models.
- Advances in computer technology have enhanced prognostic science, but there has been a lag in automated data acquisition.
- Users of ICU performance benchmarks need assistance from a nonprofit, commercial, or government entity.
- Success in European and Veterans Administration hospitals suggests that government support may be required to promote ICU mortality benchmarks.

Lessons learned are as follows:

(1) ICUs 'fight fires' (provide life-supporting therapy), but they also 'sell fire insurance', receive only technological monitoring and concentrated nursing care because of a perceived risk of needing active therapy [6].
(2) We found that 50% of ICU patients were admitted for monitoring with not only a low risk of death, but also for receiving life-supporting therapy [7].
(3) Using therapy to predict mortality leads to the conclusion – the more you do for patients the more likely they are to die. Physiological abnormalities are the critical determinant of mortality [8].
(4) It was possible to compare mortality among ICU patient groups in US hospitals and internationally [9].

THE 1980S: MODELS FOR PREDICTING ICU OUTCOMES PROLIFERATE

During the 1980s, Jack was ICU director and Bill headed the ICU research unit, but our relationship changed: Bill was a junior staff member, but my mentor in outcomes research and Doug Wagner dragged me into the world of statistics. We were encouraged by the positive reaction of the clinical community to APACHE I, but the system needed refinement, greater independence from therapy, and multiinstitutional validation.

Studies by others convinced us that we had underweighted APACHE I's measure for neurological function [10] and, at a time when data collection was almost exclusively manual, it required simplification. With government grant support, we published APACHE II in 1985 using data for 5815 ICU admissions at 13 hospitals [11]. The number of physiologic measures of severity was reduced from 34 to 12 and mortality prediction was adjusted for 44 diagnoses. By today's standards, the number of patients and variables were small, but the mainframe computer used to develop APACHE II occupied a basement and some analyses required an entire weekend. A conceptual outline of what we believed determined mortality for ICU patients during the 1980s is shown in Table 1.

APACHE II was widely used and ultimately received over 3000 citations. The need to simplify severity measurement and mortality prediction was also emphasized by the development of the Simplified Acute Physiology Score (SAPS) [12] and the Mortality Prediction Model (MPM) [13].

Lessons learned are as follows:

(1) Physiological abnormalities are directly related to mortality, but ICU admission diagnosis is also critical for accurate mortality prediction [11].
(2) Risk-adjusted mortality could be compared across ICUs in US medical centers and differences in management were evident at ICUs with superior vs. inferior performance [14].
(3) Take care with what you share. We were surprised that SAPS used the same physiological abnormalities and weights as APACHE II [12].

1988: APACHE MEDICAL SYSTEMS ESTABLISHED

Following APACHE II's publication, independent investigators reported important short-comings: lack of adjustment for patient selection, location before ICU admission, lead time bias, and concerns about the timing of data collection [15]. In addition, requests for user assistance overwhelmed our research team. We needed funding to expand the

Table 1. Determinants of hospital mortality for ICU patients: a conceptual outline from the 1980s

Information available prior to ICU treatment	
Patient factor	Predictor variables
Type of disease	ICU admission diagnosis
	Emergency vs. elective surgery
Physiologic reserve	Age
	Chronic health status
Severity of illness	Physiological abnormalities
Patient information available after treatment	
Treatment factors	Not used for prediction
Type of therapy available	
Use or application of therapy	
Timing and process of care	

APACHE database, improve accuracy, and develop software and benchmarks for resource use. Interest in further analysis of the relationship between ICU management and performance resulted in generous government and foundation grant support, but we were unable to obtain funds to support patient data collection. In 1988, it did not seem likely that a third generation of APACHE could be developed.

Fortunately, a senior radiologist at George Washington University collaborating with several venture capitalists helped our research group to found APACHE Medical Systems Inc. (AMSI). In 1988, transfer of technology and academic–commercial intellectual property agreements was unusual, although basic scientists often entered into these arrangements. During 1988–1990, AMSI was almost exclusively dedicated to collecting the APACHE III database. Academic appointments precluded Bill, Jack, and Doug from direct participation, but Betty Draper resigned from the ICU research unit to lead AMSI.

Lessons learned are as follows:

(1) Jack learned what it means to believe in your research, particularly when I discussed obtaining a second home mortgage with my wife to found AMSI.
(2) We were surprised at the willingness of corporations to commit venture capital to support AMSI and the development of APACHE III.
(3) Betty's position at AMSI provided invaluable insights to the ICU research team. She became a focal point for user feedback, which focused our research on 'real world' problems and needs.

THE 1990S: SHOULD PREDICTIVE MODELS BE SIMPLE OR COMPLEX?

APACHE III was published in 1991 and achieved greater prognostic accuracy than its predecessor, but at the expense of increased model complexity [15]. Refinements included assessment of the predictive impact of measurement timing, missing data, testing of 34 co-morbid and 19 physiological variables, nonlinear weighting (splines) of physiological variables, and expanded adjustment for diagnosis. Increased complexity was incorporated, not only to increase accuracy, but because the power of desktop computers exceeded that of basement-sized mainframes. Technological advances also made it possible to automate data collection, analyze larger databases, and use complex models. The first commercial installation of an APACHE system was at William Beaumont Hospital, Royal Oak, Michigan in 1991.

SAPS II and MPM II were also developed in the 1990s [16,17]. Their developers did not use extensive information about diagnosis and used terms such as 'simplified' and 'parsimonious' to emphasize the need to limit complexity. There were clearly divergent opinions about whether prognostic models should be simple or complex, and data collection manual or automated.

Lessons learned are as follows:

(1) Predictions using physiologic data ± 1 h of admission were not statistically different from worst values over 24 h; fewer missing values and maximum explanatory power favored 24-h values [15].
(2) Marked variations in adjusted mortality and ICU stay were found among ICUs in the APACHE III study [18].
(3) Superior ICU performance was associated with superior technology, organizational structure, and managerial practices [19]. ICUs with superior or inferior SMRs, however, could not be distinguished by ICU clinicians and organizational researchers [20].
(4) Daily mortality estimates using ICU day 1–7 data were found to have a potential role in assessing individual prognosis [21].
(5) The use of APACHE II, APACHE III, MPM II, and SAPS II outside the United States resulted in conflicting conclusions about which model most accurately predicted mortality. Only later did we learn that model recalibration was needed before using prognostic models in a different healthcare system [22].
(6) Outcome researchers recognized that mortality is overpredicted when older models are applied to more contemporary data [23].

Sales of APACHE III, media attention, and additional private investment resulted in the ability of AMSI to make a public offering in 1996. The $25 million obtained from stock sales provided the ability to improve the software system and marketing, and confirm the accuracy of APACHE III in an independent US database [24]. Our research team was no longer located at George Washington, but the internet allowed us to continue working together.

Lessons learned are as follows:

(1) Many ICU clinicians and researchers demanded that APACHE III models, software, and services be provided free.
(2) Some clinical and academic physicians hated commercialization and expressed great hostility toward APACHE.

(3) Moving from model simplification to complexity in the 1990s represented 'A bridge too far.' Computer advances supported the development and use of APACHE III, but this was not accompanied by advances in automated data collection.
(4) Founding AMSI was not profitable. The ICU research team were reimbursed for their home mortgage loans, but none of us became wealthy.

Our research was successful, but AMSI was not. Its failure was caused by the system's cost and skepticism about ICU benchmarking. Purchasing APACHE III competed with other technologies, for example, computed tomography and MRI, and required paying a coordinator because few hospitals had full automation. By 2000, AMSI had over 100 clients and 75 employees, but was nearly bankrupt. The intellectual property of AMSI was purchased by Cerner Corporation in 2002. Further insights about the history of APACHE are reported in a 2002 review by Bill Knaus [25].

2000–2010: PROGNOSTIC SYSTEMS MATURE

When I (A.A.K.) joined Cerner Corp. in 2003, my first job was to revalidate all 77 APACHE III equations. Some APACHE III equations had been revalidated, but most had not. Changes in clinical practice suggested equations might be poorly calibrated, and this was indeed the case. Although simple recalibration was possible, there was an opportunity to develop APACHE IV, expand diagnostic groups, add and refine predictor variables, and adjust for the impact of sedation on Glasgow Coma Score [26,27]. We have recently described details about the evolution and capabilities of APACHE IV [28].

Project IMPACT, a database aimed at describing and measuring the care of ICU patients, was developed by the Society of Critical Care Medicine (SCCM) in 1996 [29]. Initially, project IMPACT used MPM II, SAPS II, and APACHE II, but later focused on MPM II to provide benchmarks for mortality and length of stay using weighted hospital days [30]. SCCM used commercial entities to develop software and Tri-Analytics Inc. to provide analyses and reports. In 2004, project IMPACT was sold to Cerner Corporation; and project IMPACT data were provided for the development and validation of MPM_{0}-III [31]. In 2005, European researchers developed SAPS 3 [32]. These contemporary models continued to differ fundamentally; APACHE IV remained complex, whereas MPMo-III and SAPS 3 emphasized simplicity.

Benchmarks using contemporary mortality and length of stay models have proven useful for assessing ICU performance for patient groups [1,4]. But predicting mortality and ICU stay for individual patients requires data for each ICU day and complex modeling, and is subject to misuse. The use of APACHE IV data from ICU day 5 has improved predictive accuracy for patients with prolonged ICU stays [33], but no contemporary model has proven suitable for predicting individual patient outcomes or making end-of-life decisions.

Lessons learned are as follows:

(1) Similarly to our experience with APACHE, SCCM found that the infrastructure for data acquisition, analysis, and reporting required commercial support.
(2) Model revalidation corrects some but not all time-related mortality overpredictions. Decreases in mortality over time are associated with improved disease-specific therapy [34▪].
(3) The electronic infrastructure for collecting data for outcome assessment was available, but not widely employed in critical care [35▪].
(4) The cost of health information technology remains high and hospital expenditures for systems to benchmark outcomes continue to fall behind spending for other technologies.

2011–PRESENT: RISK ADJUSTMENT IN THE PUBLIC REALM

The use of APACHE as a national core measure of ICU quality in the United States was recently considered by The Joint Commission (TJC), but not acted upon because of insufficient resources for implementation. In contrast, recalibrated APACHE mortality and ICU length of stay models were recently adopted by the Netherlands' National Intensive Care Evaluation program for comparing Dutch ICUs [36▪,37]. The Australia and New Zealand Intensive Care Society uses a recalibrated version of APACHE III for comparative reporting [38]. In the UK, the Intensive Care National Audit and Research Centre uses a simplified model that includes the best elements of existing models and the characteristics of their ICU population for comparative data reports and research [39]. Based on the successful use of SMRs for quality measurement internationally [1] and by the US Veterans Administration [2], government support or intervention might be required to implement quality assessment using risk-adjusted ICU outcomes.

Starting in the early 2000s, there has been an increased effort to make more hospital data electronically available. This has been expedited by

Center for Medicare Services (CMS) mandates that tie payments to reaching specified levels of electronic data capture. Further, a scheme of reimbursing hospitals resultant on outcomes measures such as mortality, length of stay, and 30-day readmission has begun. Unfortunately, payments to hospitals for some of these outcomes are withheld based on unadjusted measures. This could unfairly penalize teaching hospitals in urban centers that traditionally treat severely ill patients.

In order to fairly compare hospitals, outcomes need to be risk-adjusted. This was shown recently for ICU readmissions [40▪]. Although mortality and ICU stay were worse for ICUs in the highest tertile of readmission percentage, these differences disappeared once outcomes were risk-adjusted using APACHE IV [41▪▪]. CMS recently initiated a study group to investigate the creation of a model to assess case-mix-adjusted ICU readmission rates.

THE 'BIG DATA' EXPLOSION

The proliferation of various 'OMICs' (genomics, proteomics, metabolomics, and so on: molecular aspects of biologic processes) has resulted in terabytes of data being available for modeling; the so-called 'Big Data' phenomenon. However, the use of such information necessitates that it be available electronically. Further, the time period between when a sample is taken and results are generated must be shortened considerably to be used in predictive equations. These requirements entail that a sophisticated, interconnected electronic medical record system be ubiquitous throughout a hospital.

As shown in Table 2, current knowledge about the determinants of hospital mortality and resource use for ICU patients has markedly expanded. The incorporation of these concepts in addition to the concept of using 'Big Data' for predictive models runs counter to attempts to provide simplified models. The latter have been shown to be less effective for predicting mortality for ICU patient groups [42▪▪] and for benchmarking ICU performance [43]. It is inconceivable that simple models could be effectively used for predicting individual patients' outcomes.

There have been legitimate concerns about solely using SMRs for quality assessment [1,44].

Table 2. Determinants of hospital mortality and resource use for ICU patients: A conceptual outline from 2014

Patient information available before ICU treatment	
Patient factors	Predictor variables
Type of disease	Specific ICU admission diagnosis
	Emergency vs. elective surgery
	Diagnosis specific risk factors (CABG and AMI)
Physiologic reserve	Age
	Specific comorbid conditions
Severity of illness	APS
Patient information available after treatment	
Treatment factors	Predictor variables
Treatment before ICU admission	Location and length of stay before admission
	Life support before ICU admission
Treatment status at ICU admission	Intensive monitoring vs. active therapy
	Life-support that influences physiological measures; includes mechanical ventilation, vasoactive drug infusion, sedation, paralysis, DNR orders, treatment limitations
Response to treatment	Daily physiological measures (APS)
	Point in time measures (e.g., day 5 APS, discharge APS)
Institution-based adjustments	
Healthcare system	Calibration for country or region
Hospital characteristics	Hospital bed size, teaching status
Duration of therapy	Variations in duration of ICU and hospital stay
	Discharge to postacute care facility
ICU-level differences	Diagnostic case-mix
	Severity of illness
	Frequency of mechanical ventilation
	Frequency of transfer from other hospitals

AMI, acute myocardial infarction; APS, Acute Physiology Score; CABG, coronary artery bypass graft; DNR, do-not-resuscitate.

Optimizing the use of 'Big Data' for outcome predictions also requires rethinking of how data are stored and analyzed. There is a strong movement toward abandoning typical database management systems that use a relational design and adopting the more flexible Hadoop architecture [45]. There is also much discussion about whether models should be developed using traditional statistical techniques that rely on linear or logistic regression. If nonlinear machine learning methods such as random forests [46] are used to generate predictions, then parallel processing needs to be considered.

CONCLUSION

Models for predicting ICU outcomes have already gone through numerous cycles. The next generation of predictive models should be more accurate and timely, and include measures of resource and financial use.

Acknowledgements

This review was supported by Cerner Corporation, Kansas City, Missouri, USA.

Conflicts of interest

Dr Zimmerman is a consultant and receives research support and speaking fees from Cerner Corporation. Dr Kramer is an employee and stock holder of Cerner Corporation.

REFERENCES AND RECOMMENDED READING

Papers of particular interest, published within the annual period of review, have been highlighted as:

- ■ of special interest
- ■■ of outstanding interest

1. Flaatten H. The present use of quality indicators in the intensive care unit. Acta Anaesthesiol Scand 2012; 56:1078–1083.
2. Render ML, Freyberg RW, Hasselbeck R, *et al.* Infrastructure for quality transformation: measurement and reporting in veterans administration intensive care units. BMJ Qual Saf 2011; 20:498–507.
3. Breslow MJ, Badawi O. Severity scoring in the critically ill. Part 1: Interpretation and accuracy of outcome prediction scoring systems. Chest 2012; 141:245–252.
4. Breslow MJ, Badawi O. Severity scoring in the critically ill. Part 2: maximizing value from outcome prediction scoring systems. Chest 2012; 141:518–527.
5. Knaus WA, Zimmerman JE, Wagner DP, *et al.* APACHE – Acute Physiology and Chronic Health Evaluation: a physiologically based classification system. Crit Care Med 1981; 9:591–597.
6. Knaus WA, Wagner DP, Draper EA, *et al.* The range of intensive care services today. JAMA 1981; 246:2711–2716.
7. Wagner DP, Knaus WA, Draper EA, *et al.* Identification of low-risk monitor patients within a medical-surgical intensive care unit. Med Care 1983; 21:425–434.
8. Scheffler RM, Knaus WA, Wagner DP, *et al.* Severity of illness and the relationship between intensive care and survival. Am J Public Health 1982; 72:449–454.
9. Knaus WA, Le Gall JR, Wagner DP, *et al.* A comparison of intensive care in the U.S.A and France. Lancet 1982; ii:642–646.
10. Teres D, Brown RB, Lemeshow S. Predicting mortality of intensive care unit patients: the importance of coma. Crit Care Med 1982; 10:86–95.
11. Knaus WA, Draper EA, Wagner DP, *et al.* APACHE II: a severity of disease classification system. Crit Care Med 1985; 13:818–829.
12. Le Gall JR, Loirat P, Alperovitch A, *et al.* A simplified acute physiology score for ICU patients. Crit Care Med 1984; 12:975–977.
13. Lemeshow S, Teres D, Pastides H, *et al.* A method for predicting survival and mortality of ICU patients using objectively derived weights. Crit Care Med 1985; 13:519–525.
14. Knaus WA, Draper EA, Wagner DP, *et al.* Evaluation of outcome from intensive care in major medical centers. Ann Intern Med 1986; 104:410–418.
15. 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.
16. Le Gall JR, Lemeshow S, Saulnier F. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. JAMA 1993; 270:2957–2963.
17. Lemeshow S, Teres D, Klar J, *et al.* Mortality Probability Models (MPM II) based on an international cohort of intensive care unit patients. JAMA 1993; 270:2478–2486.
18. Knaus WA, Wagner DP, Zimmerman JE, *et al.* Variations in mortality and length of stay in intensive care units. Ann Intern Med 1993; 118:753–761.
19. Shortell SM, Zimmerman JE, Rousseau DM, *et al.* The performance of intensive care units: does good management make a difference. Med Care 1994; 32:508–525.
20. Zimmerman JE, Shortell SM, Rousseau DM, *et al.* Improving intensive care: observations based on organizational case studies in nine intensive care units: a prospective multicenter study. Crit Care Med 1993; 21:1443–1451.
21. Wagner DP, Knaus WA, Harrell FE, *et al.* Daily prognostic estimates for critically ill adults in intensive care units: Results from a prospective, multicenter, inception cohort analysis. Crit Care Med 1994; 22:1359–1372.
22. Harrison DA, Brady AR, Parry GJ. Recalibration of risk prediction models in a large multicenter cohort of admissions to adult, general critical care units in the United Kingdom. Crit Care Med 2006; 34:1378–1388.
23. Kramer AA. Predictive mortality models are not like fine wine. Crit Care 2005; 9:636–637.
24. Zimmerman JE, Wagner DP, Draper EA, *et al.* Evaluation of Acute Physiology and Chronic Health Evaluation III predictions of hospital mortality in an independent database. Crit Care Med 1998; 26:1317–1326.
25. Knaus WA. APACHE 1978-2001: The development of a quality assurance system based on prognosis. Milestones and personal reflections. Arch Surg 2002; 137:37–41.
26. Zimmerman JE, Kramer AA, McNair DS, *et al.* Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patents. Crit Care Med 2006; 34:1297–1310.
27. Zimmerman JE, Kramer AA, McNair DS, *et al.* Intensive care unit length of stay: Benchmarking based on Acute Physiology and Chronic Health Evaluation (APACHE) IV. Crit Care Med 2006; 34:2517–2529.
28. Zimmerman JE, Kramer AA. Outcome prediction in critical care: the Acute Physiology and Chronic Evaluation models. Curr Opin Crit Care 2008; 14:491–497.
29. Cook SF, Visscher WA, Hobbs CL, *et al.* Project IMPACT: results from a pilot validity study of a new observational database. Crit Care Med 2002; 30:2765–2770.
30. Higgins TL, Kramer AA, Nathanson BH, *et al.* Prospective validation of the intensive care unit admission Mortality Probability Model (MPM_O-III). Crit Care Med 2009; 37:1619–1623.
31. Nathanson BH, Higgins TL, Teres D, *et al.* A revised method to assess intensive care unit clinical performance and resource utilization. Crit Care Med 2007; 35:1853–1862.
32. Moreno RP, Metnitz PGH, Almeida E, *et al.* 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.
33. Kramer AA, Zimmerman JE. A predictive model for the early identification of patients at risk for a prolonged intensive care unit length of stay. BMC Med Inform Decis Mak 2010; 10:27.
34. ■ Zimmerman JE, Kramer AA, Knaus WA. Changes in hospital mortality for United States intensive care unit admissions from 1988 to 2012. Crit Care 2013; 17:R81.
Analysis of aggregate and disease-specific decreases in mortality over the last 24 years among ICU patients included in APACHE databases.
35. ■ Chen LM, Kennedy EH, Sales A, *et al.* The use of health IT for higher-value critical care. N Engl J Med 2013; 368:594–597.
Perspective about how health information technology can be used to provide decision support, and improve patient triage and outcome assessment.
36. ■ Brinkman S, Abu-Hanna A, de Jonge E, *et al.* Prediction of long-term mortality in ICU patients: model validation and assessing the effect of using in-hospital versus long-term mortality on benchmarking. Intensive Care Med 2013; 39:1925–1931.
Customized APACHE IV mortality models are used to demonstrate the influence of discharge policies on case-mix-adjusted mortality.
37. Brandenburg R, Brinkman S, de Keizer NF, *et al.* In-hospital mortality and long-term survival of patients with acute intoxication admitted to the ICU. Crit Care Med 2014; 42:1471–1479.

38. Moran JL, Bristow P, Solomon PJ, *et al.* Mortality and length-of-stay outcomes, 1993–2003, in the binational Australian and New Zealand intensive care adult patient database. Crit Care Med 2008; 36:46–61.

39. Harrison DA, Rowan KM. Outcome prediction in critical care: the ICNARC model. Curr Opin Crit Care 2008; 14:506–512.

40. ▪ Kramer AA, Higgins TL, Zimmerman JE. Intensive care unit readmissions in U.S. hospitals: patient characteristics, risk factors, and outcomes. Crit Care Med 2012; 40:3–10.
Analysis of the risk factors and outcomes of patients readmitted to ICU. These patients have greater severity and complexity of illness, increased mortality, and greater ICU and hospital lengths of stay.

41. ▪▪ Kramer AA, Higgins TL, Zimmerman JE. The association between intensive care unit readmission rate and patient outcomes. Crit Care Med 2013; 41:24–33.
ICUs with readmission rates ranging from 1.2 to 14.5% had similar standardized mortality and ICU length of stay after adjustment using APACHE IV. ICU readmission rate should be used as a quality measure only if case-mix adjustment is taken into account.

42. ▪▪ 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.
APACHE mortality predictions offered the best discrimination, calibration, and accuracy compared with MPM_0-III and a National Quality Forum endorsed model for identical patients in a multi-institutional dataset.

43. Kramer AA, Higgins TL, Zimmerman JE. Ranking ICU performance: does the benchmarking tool matter? Am J Respir Crit Care Med 2014; 189:A2526.

44. Nassar AP, Mocelin AO, Nunes ALP, *et al.* Caution when using prognostic models: a prospective comparison of 3 recent prognostic models. J Crit Care 2012; 27:423; e1–7.

45. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Health Inf Sci and Systems 2014; 2:1–10.

46. Breiman L. 'Statistical modeling: the two cultures'. Statistical Science 2001; 16:199–231.

Why try to predict ICU outcomes?

G. Sarah Power and David A. Harrison

Purpose of review
To describe why the prediction of ICU outcomes is essential to underpin critical care quality improvement programmes.

Recent findings
Recent literature demonstrates that risk-adjusted mortality is a widely used and well-accepted quality indicator for benchmarking ICU performance. Ongoing research continues to address the best ways to present the results of benchmarking through either direct comparison among institutions (e.g., by funnel plots) or indirect comparison against the risk predictions from a risk model (e.g., by process control charts). There is also ongoing research and debate regarding event-based outcomes (e.g., hospital mortality) versus time-based outcomes (e.g., 30-day mortality). Beyond benchmarking, ICU outcome prediction models have a role in risk adjustment and risk stratification in randomized controlled trials, and adjusting for confounding in nonrandomized, observational research. Recent examples include comparing risk-adjusted outcomes according to 'capacity strain' on the ICU and extending propensity matching methods to evaluate outcomes of patients managed with a pulmonary artery catheter, among others. Risk models may have a role in communicating risk, but their utility for individual patient decision-making is limited.

Summary
Risk-adjusted mortality has strong support from the critical care community as a quality indicator for benchmarking ICU performance but is dependent on up-to-date, accurate risk models. ICU outcome prediction can also contribute to both randomized and nonrandomized research and potentially contribute to individual patient management, although generic risk models should not be used to guide individual treatment decisions.

Keywords
benchmarking, quality improvement, risk adjustment, severity of illness index, statistical models

INTRODUCTION

Quality improvement in healthcare requires assessment of the structure, processes and outcomes of care. In this review, we describe this framework in more detail, with particular reference to critical care, and describe why predicting ICU outcomes is essential to underpin critical care quality improvement programmes. We review the recent literature regarding quality improvement in critical care, focusing on the use of risk-adjusted mortality as a quality indicator. Finally, we describe other uses for outcome prediction in critical care, with reference to recently published papers.

A FRAMEWORK FOR QUALITY IMPROVEMENT

Healthcare professionals agree that the quality of care received by a patient should be to the highest possible standard; however, it has long been recognized that variations in healthcare practice exist [1]. Coupled with this is a pressure for healthcare providers to reduce resource consumption [2]. In 1999, the US Institute of Medicine published To Err Is Human: Building a Safer Health System, which concluded that thousands of Americans experience preventable medical error each year [3], and a subsequent 2001 report proposed a comprehensive strategy to improve the quality and delivery of care in the United States [4]. That said, quality improvement was by no means a new concept at the start of the new millennium. Some 10 years earlier, the same organization had defined quality as 'the degree to which health services for individuals and

Intensive Care National Audit & Research Centre (ICNARC), London, United Kingdom

Correspondence to Dr David A. Harrison, Senior Statistician, ICNARC, Napier House, 24 High Holborn, London WC1V 6AZ, UK. Tel: +44 20 7831 6878; fax: +44 20 7831 6879; e-mail: david.harrison@icnarc.org

Curr Opin Crit Care 2014, 20:544–549
DOI:10.1097/MCC.0000000000000136

KEY POINTS

- Quality improvement in healthcare requires assessment of the structure, processes and outcomes of care.
- Comparative audit of outcomes requires risk adjustment to allow fair comparison of institutions.
- Risk-adjusted mortality is widely used and well accepted as a quality indicator for benchmarking ICU performance.
- Beyond benchmarking, ICU outcome prediction models can contribute to randomized and nonrandomized research and patient management.

populations increase the likelihood of desired health outcomes and are consistent with current professional knowledge' [5]. Much earlier, in 1917, the American College of Surgeons proposed 'minimum standards' when establishing its Hospital Standardization Programme, which included keeping medical records that contain the history, physical examination and laboratory results and limiting staff membership to well-educated, competent and licensed physicians and surgeons [6].

Quality improvement in healthcare has clearly come a long way over the last century. The most widely used framework for assessing the quality of healthcare is that proposed by Donabedian. In his 1988 *JAMA* report, he describes three distinct categories from which inferences about the quality of care can be drawn: structure, process and outcome [7]. Structure describes the attributes of the setting within which care occurs. Process denotes what is actually done in giving and receiving care. Outcome is defined as the effects of care on the health status of patients and populations. For example, in a critical care setting, indicators of structure may include physical design of the ICU and staffing levels, indicators of process may include adherence to lung protective ventilation strategies and appropriate use of stress ulcer prophylaxis, and indicators of outcome may include mortality and infection rates.

The domains of structure and process can be assessed within an individual institution. Indicators of structure are assessed against professional standards, regulations and recommendations. They relate to the institution rather than the patient, and therefore require only periodic assessment. Indicators of process are assessed against national or international clinical guidelines, based ideally on high-quality evidence but often on expert consensus. A process audit involves identifying all patients in a time period who were eligible for a particular pathway or protocol, establishing whether the pathway was correctly followed, and identifying areas for improvement. The gold standard is 100% compliance and there is no need to compare performance against other institutions.

Assessing outcomes for a single institution is not as straightforward and requires comparison against other institutions – comparative audit – to put the outcome of the particular institution in context and to enable benchmarking. However, the quality of care is only one of many factors that will contribute to a patient's outcome and, if crude outcomes were to be compared between institutions, any effect of quality would likely be overwhelmed by variation in the patient demographics, underlying health status, acute conditions and severity of the acute illness (collectively termed case mix). When comparing outcomes between institutions, it is therefore essential to take the differing case mix of the institutions into account to be able to make fair comparisons.

ICU OUTCOME PREDICTION FOR QUALITY IMPROVEMENT

The most widely used ICU outcome measure is mortality, as it is patient-centered, objective and easily measured up to the point of hospital discharge [8]. In a recent review of ICU quality indicators, Flaatten [9] identified eight sets of ICU quality indicators developed at a national level. Risk-adjusted mortality was the most frequently included indicator, being included in six of the eight quality indicator sets (from Austria, India, the Netherlands, Scotland, Sweden and the United Kingdom). The two quality indicator sets that did not include risk-adjusted mortality were from Germany, where it was not considered in the development of the quality indicator set, as it was already included in a core national dataset, and Spain (http://www.semicyuc.org/temas/calidad), where it was included in a wider list of 120 indicators but not their 20 'fundamental' indicators. Using a modified Delphi process with a group of 18 nominated experts from nine countries, a European Society of Intensive Care Medicine Task Force on Quality and Safety identified nine quality indicators for intensive care medicine [10]. Risk-adjusted mortality was one of seven indicators reaching 100% consensus in the final round of the Delphi process.

The purpose of risk prediction models in intensive care is to take physiological data from early in the critical illness, ideally prior to intervention but in practice often over the first 24 h following admission, coupled with other patient risk factors such as age and diagnostic coding/reason for admission to predict the risk of hospital mortality for each

patient. These predicted risks can be used to evaluate the outcome of one institution compared with others either directly, by comparing risk-adjusted outcomes between the institutions, or indirectly, by comparing outcomes for the single institution against those predicted by the model.

Within critical care, the most commonly used and agreed upon measure for direct comparison is standardized mortality ratio (SMR). The SMR is calculated for each institution by summing the total number of deaths over a given time period and dividing this by the expected number of deaths as predicted by the risk model, calculated by summing the predicted risks over the same time period. Perfect agreement leads to a SMR of one, although there will of course be variation around this value. Funnel plots are an application of direct comparison; in this case 'acceptable' limits of variation from one will depend on the SMRs of all ICUs being examined (Fig. 1) [11]. The advantage of funnel plots is that they take into account the chance variation in the outcome and therefore avoid institutions being ranked as the best or worst based on imprecise estimates. However, care must be taken when using funnel plots to identify potential outliers [12▪▪]. The power of the funnel plot to detect whether an institution is an outlier is very dependent on the number of events/patients within that institution. When the number of events is small, the probability of an institution being identified as a potential outlier, even when there is true poor performance, would be low. Conversely, for a large institution, even a very small increase in deaths from that predicted by the risk model may result in the institution being identified as a potential outlier. Any inferences about quality of care are dependent on the risk model being accurate and up-to-date to avoid the majority of participants being considered 'better than average.' Recently, Tran *et al.* [13] showed that the EuroSCORE, developed in the mid-1990s to predict mortality after cardiac surgery, 'significantly underestimated' the risk-adjusted mortality for all surgeons prior to recalibration. Within a critical care setting, Harrison *et al.* [14] demonstrated the importance of recalibration as well as the need to use models that are calibrated to the population of interest.

Indirect comparison compares the observed outcomes for a single institution with the expected outcomes as predicted by the risk model, that is, indirectly comparing this institution's outcomes against those of the institutions used to develop or recalibrate the risk model. Process control charts can be used to make comparisons of the observed and expected outcomes within an institution. Koetsier *et al.* [15] recently reviewed a number of alternative process control methods in the context of simulated ICU data. They found that exponentially weighted moving average (EWMA) charts were the most sensitive to detect an upward shift in the mortality in an institution. EWMA charts are also intuitively understandable, as they plot the average

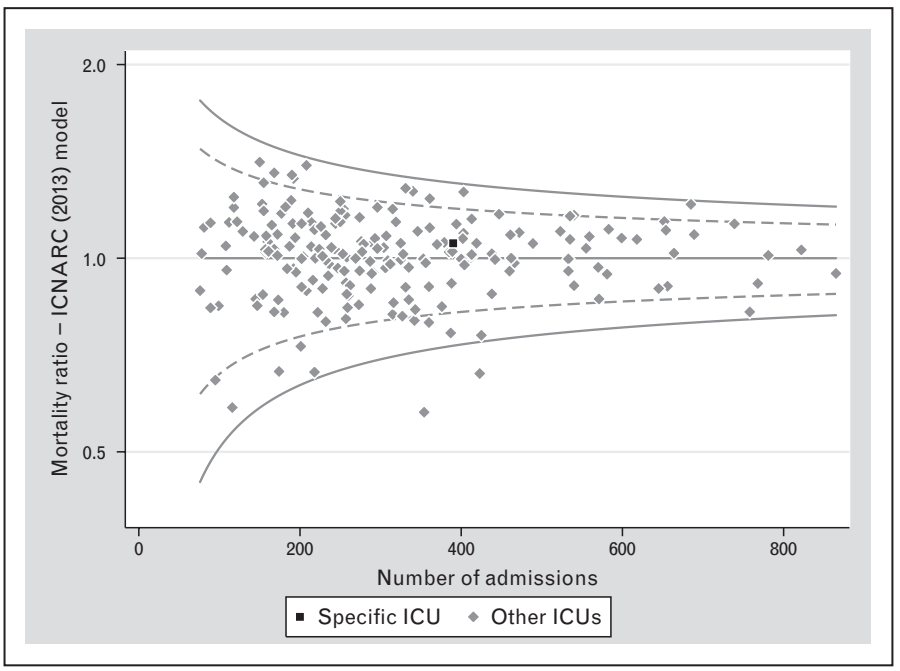

FIGURE 1. Example of a funnel plot for a fictitious ICU, based on data from the Intensive Care National Audit & Research Centre (ICNARC) Case Mix Programme with risk adjustment using the ICNARC model, 2013 recalibration.

observed and predicted mortality rate, updated after each new patient (Fig. 2). The indirect comparison approach may seem appealing, as it requires only data from a single institution and the risk model. However, data from multiple institutions are still required to build or recalibrate a risk model and any inferences about quality of care are even more dependent on the risk model being accurate and up-to-date to avoid false reassurance from comparing a single institution against an out-of-date model.

Regardless of the method applied, it is of great importance to be able to identify ICUs with higher than expected risk-adjusted mortality, to enable investigations into the reasons why to be undertaken. These investigations should be conducted with care and consider the quality of the data as well as the appropriateness of the risk model (e.g. does it include all the important potential confounders) and its statistical performance.

Within critical care, the concept of predicting outcome has been in use for over three decades and all of the major models predict event-based outcomes (mortality at discharge from hospital). It has been argued that time-based outcomes (such as 30-day mortality) are less biased than event-based outcomes [16▪]; however, following up ICU patients to a specific time point, which may often be after discharge from hospital, is burdensome and may not be feasible to undertake routinely for the purpose of auditing ICUs. Many countries are now establishing more robust systems for patient identification across routine healthcare datasets permitting linkage to death registrations, which would address this issue. However, it would be necessary to ensure any risk models were recalibrated to the new outcome – otherwise, it is the equivalent of allowing the predicted number of deaths from the model to stay the same, while allowing the observed number of deaths to change in line with the new outcome being used. Brinkman *et al.* [17▪▪] recently compared event-based and time-based outcomes using the Acute Physiology And Chronic Health Evaluation (APACHE) IV model in ICUs in The Netherlands and recalibrated the model to each of the outcomes presented prior to undertaking any comparisons between the model performances. They found that both SMR and SMR rank were influenced by which outcome was used and recommended that SMRs based on fixed time points are preferable for use as a quality indicator.

OTHER USES FOR ICU OUTCOME PREDICTION

In addition to comparative audit and benchmarking, there are a number of other uses for ICU outcome predictions to support research and patient management.

In a research context, risk models can be used in both randomized and nonrandomized (observational) studies. Randomized controlled trials (RCTs)

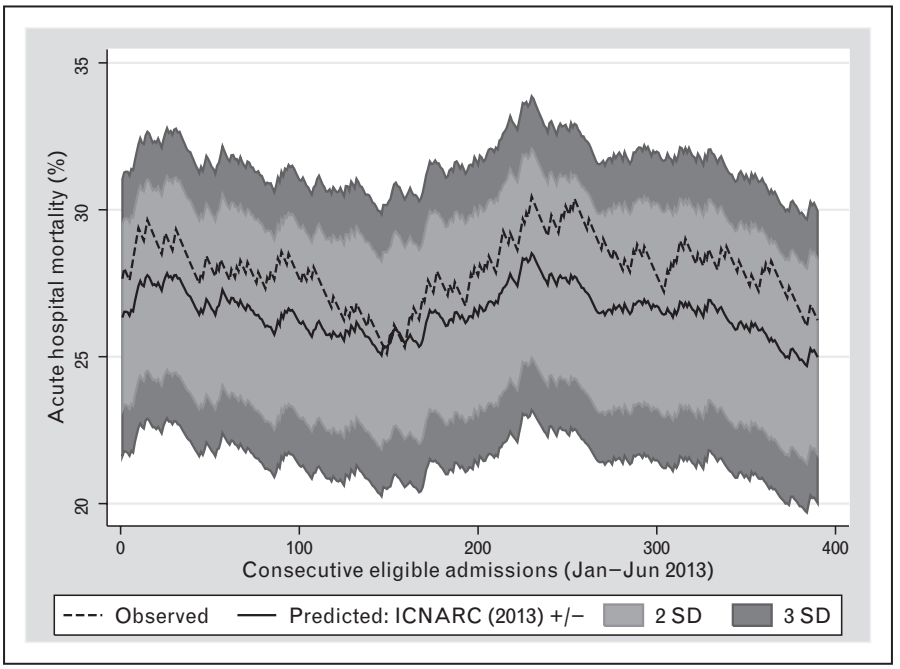

FIGURE 2. Example of an exponentially weighted moving average plot for a fictitious ICU, based on data from the Intensive Care National Audit & Research Centre (ICNARC) Case Mix Programme risk adjustment using the ICNARC model, 2013 recalibration.

usually report unadjusted patient outcomes for the different treatment groups. Although randomization is the gold standard when one wishes to investigate a causal relationship between a treatment and an outcome, as it eliminates bias in treatment allocation; trials often adjust for important predictors of outcome to correct for chance imbalances between treatment groups at baseline [18]. Using simulated datasets, Hernandez *et al.* [19] describe how covariate adjustment for a-priori specified predictors of outcome can increase the statistical power of a trial, thus reducing the sample size requirements. This is supported by Roozenbeek *et al.* [20] using data from the International Mission for Prognosis and Analysis of Clinical Trials in traumatic brain injury (IMPACT) database of RCTs and observational studies in traumatic brain injury (TBI). Nevertheless, there is not a general consensus on the best approach. Turner *et al.* [21] caution against using this approach in the planning of RCTs, but do conclude that 'moderate gains in power may be obtained using covariate adjustment from logistic regression in heterogeneous conditions such as TBI'.

When subgroup analyses of a positive RCT are unrevealing, such findings are commonly used to argue that the treatment's benefits apply to the entire study population; however, it has been contended that such analyses are often limited by low statistical power [22]. Multivariable risk-stratified analyses have been investigated as an alternative to conventional subgroup analyses and conclude that, although conventional subgroup analyses can be useful under some circumstances, clinical trial reporting should include a multivariable risk-stratified analysis when an adequate externally developed risk prediction tool is available. Kent and Hayward [23] argue that risk stratification in the reporting of RCTs is required to aid clinicians in their individual patient treatment decisions, given the 'average' patient does not present in the real world. This is further described in a second article by Kent *et al.* [24], which goes on to explain 'why risk-stratified analyses should be performed whenever feasible' and in which a framework is provided to prioritize the analysis and reporting of risk-stratified subgroups.

Where randomization is not possible for practical or ethical reasons, risk adjustment allows conclusions to be drawn from observational data. For example, in a recent article using data from the US Project IMPACT database, Wagner *et al.* [25▪] compared risk-adjusted outcomes for patients discharged from critical care according to the 'capacity strain' on the ICU. They used risk predictions from the Mortality Probability Model at ICU admission Version III (MPM0-III) risk model both to contribute to the risk adjustment at the patient level and also to define one of the measures of strain – the average predicted risk of the other patients in the ICU.

Risk models can also contribute to more complex statistical methods to adjust for confounding in observational research. For example, Sekhon and Grieve [26] recently published a new method (termed Genetic Matching) as an extension to propensity score matching. They applied their method to a case study of outcomes from ICU patients managed with a pulmonary artery catheter, comparing the results with a contemporaneous RCT. It is important to remember, however, that such advanced statistical methods only improve the ability to adjust for measured confounders and such studies remain subject to potential bias from unmeasured confounders.

The use of ICU outcome prediction in the management of individual patients is more contentious. Risk models and severity scores may provide a tool for communication between healthcare professionals [27], and may assist clinicians in providing objective estimates of likely outcomes to a patient's family [28]. However, it is generally agreed that they are not appropriate to rely on for decisions regarding specific treatments [8,29] or for limitation of life-sustaining therapy [8,30].

CONCLUSION

Risk-adjusted mortality has strong support from the critical care community as a quality indicator for benchmarking ICU performance. The utility of risk-adjusted mortality as a quality indicator is dependent on up-to-date, accurate risk models, and ongoing research continues to address the best ways to measure and report this outcome. ICU outcome prediction can also contribute to both randomized and nonrandomized research and potentially support patient management, although generic risk models should not be used to guide individual treatment decisions.

Acknowledgements

None.

Conflicts of interest

There are no conflicts of interest.

REFERENCES AND RECOMMENDED READING

Papers of particular interest, published within the annual period of review, have been highlighted as:

- ▪ of special interest
- ▪▪ of outstanding interest

1. Conry MC, Humphries N, Morgan K, *et al.* A 10 year (2000–2010) systematic review of interventions to improve quality of care in hospitals. BMC Health Serv Res 2012; 12:275.

2. Rotondi AJ, Sirio CA, Angus DC, Pinsky MR. A new conceptual framework for ICU performance appraisal and improvement. J Crit Care 2002; 17:16–28.
3. Institute of Medicine. To err is human: building a safer health system. Washington, DC: The National Academic Press; 2000; pp. 1–287.
4. Institute of Medicine. Crossing the quality chasm: a new health system for the 21st century. Washington, DC: The National Academic Press; 2001; pp. 1–337.
5. Institute of Medicine. Medicare: a strategy for quality assurance, volume I. Washington, DC: The National Academic Press; 1990; pp. 1–441.
6. Luce JM, Bindman A B, Lee P R. A brief history of healthcare quality assessment and improvement in the United States. West J Med 1994; 160:263–268.
7. Donabedian A. The quality of care. How can it be assessed? JAMA 1988; 260:1743–1748.
8. Higgins TL. Quantifying risk and benchmarking performance in the adult intensive care unit. J Intensive Care Med 2007; 22:141–156.
9. Flaatten H. The present use of quality indicators in the intensive care unit. Acta Anaesthesiol Scand 2012; 56:1078–1083.
10. Rhodes A, Moreno RP, Azoulay E, *et al.* Prospectively defined indicators to improve the safety and quality of care for critically ill patients: a report from the Task Force on Safety and Quality of the European Society of Intensive Care Medicine (ESICM). Intensive Care Med 2012; 38:598–605.
11. Spiegelhalter DJ. Funnel plots for comparing institutional performance. Stat Med 2005; 24:1185–1202.
12. ■■ Seaton SE, Barker L, Lingsma HF, *et al.* What is the probability of detecting poorly performing hospitals using funnel plots? BMJ Qual Saf 2013; 22:870–876.

This study explores the strengths and limitations of the funnel plot methodology, commonly used for benchmarking ICU performance.

13. Tran DT, Dupuis JY, Mesana T, *et al.* Comparison of the EuroSCORE and Cardiac Anesthesia Risk Evaluation (CORE) score for risk-adjusted mortality analysis in cardiac surgery. Eur J Cardiothorac Surg 2012; 41:307–313.
14. Harrison DA, Brady AR, Parry GJ, *et al.* Recalibration of risk prediction models in a large multicenter cohort of admissions to adult, general critical care unit in the United Kingdom. Crit Care Med 2006; 34:1378–1388.
15. Koetsier A, de Keizer NF, de Jonge E, *et al.* Performance of risk-adjusted control charts to monitor in-hospital mortality of intensive care unit patients: a simulation study. Crit Care Med 2012; 40:1799–1807.
16. ■ Reineck LA, Pike F, Le TQ, *et al.* Hospital factors associated with discharge bias in ICU performance measurement. Crit Care Med 2014; 42:1055–1064.

This study identifies potential sources of bias from using hospital mortality, compared with 30-day mortality, as an outcome for ICU performance assessment.

17. ■■ Brinkman S, Abu-Hanna A, de Jonge E, de Keizer NF. Prediction of long-term mortality in ICU patients: model validation and assessing the effect of using in-hospital versus long-term mortality on benchmarking. Intensive Care Med 2013; 39:1925–1931.

This study recalibrates the APACHE IV model to mortality at 1, 3 and 6 months following ICU admission, and compares benchmarking on these outcomes with benchmarking on hospital mortality, demonstrating substantial differences.

18. Steyerberg EW, Bossuyt PM, Lee KL. Clinical trials in acute myocardial infarction: should we adjust for baseline characteristics? Am Heart J 2000; 139:745–751.
19. Hernandez AV, Steyerberg EW, Habbema JD. Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. J Clin Epidemiol 2004; 57:454–460.
20. Roozenbeek B, Maas AI, Lingsma HF, *et al.* Baseline characteristics and statistical power in randomized controlled trials: selection, prognostic targeting, or covariate adjustment? Crit Care Med 2009; 37:2683–2690.
21. Turner EL, Perel P, Clayton T, *et al.* Covariate adjustment increased power in randomized controlled trials: an example in traumatic brain injury. J Clin Epidemiol 2012; 65:474–481.
22. Hayward RA, Kent DM, Vijan S, Hofer TP. Multivariable risk prediction can greatly enhance the statistical power of clinical trial subgroup analysis. BMC Med Res Methodol 2006; 6:18.
23. Kent DM, Hayward RA. Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification. JAMA 2007; 298:1209–1212.
24. Kent DM, Rothwell PM, Ioannidis JP, *et al.* Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. Trials 2010; 11:85.
25. ■ Wagner J, Gabler NB, Ratcliffe SJ, *et al.* Outcomes among patients discharged from busy intensive care units. Ann Intern Med 2013; 159:447–455.

This study is a good example of an observational study making use of risk models for both risk adjustment and defining an exposure variable.

26. Sekhon JS, Grieve RD. A matching method for improving covariate balance in cost-effectiveness analyses. Health Econ 2012; 21:695–714.
27. Gunning K, Rowan K. ABC of intensive care: outcome data and scoring systems. BMJ 1999; 319:241–244.
28. Knaus WA, Harrell FE Jr, Lynn J, *et al.* The SUPPORT prognostic model. Objective estimates of survival for seriously ill hospitalized adults. Ann Intern Med 1995; 122:191–203.
29. Vincent JL, Opal SM, Marshall JC. Ten reasons why we should NOT use severity scores as entry criteria for clinical trials or in our treatment decisions. Crit Care Med 2010; 38:283–287.
30. Skrobik Y, Kavanagh BP. Scoring systems for the critically ill: use, misuse and abuse. Can J Anaesth 2006; 53:432–436.

ICU severity of illness scores: APACHE, SAPS and MPM

Jorge I.F. Salluh[a,b] *and Márcio Soares*[a,b]

Purpose of review
This review aims to evaluate the latest versions of the Acute Physiology and Chronic Health Evaluation, Simplified Acute Physiology Score and Mortality Probability Model scores, make comparisons and describe their strengths and limitations. Additionally, we provide critical analysis and recommendations for the use of these scoring systems in different scenarios.

Recent findings
The last generation of ICU scoring systems (Acute Physiology and Chronic Health Evaluation IV, Mortality Probability Model 0–III (MPM0-III) and Simplified Acute Physiology Score 3) was widely validated in different regions of the world and in distinct settings comprising general ICU patients as well as specific subgroups such as critically ill cancer patients, cardiovascular, surgical, acute kidney injury requiring renal replacement therapy and those in need of extra-corporeal membrane oxygen. Conflicting results are reported, and in general the scores presented a good discrimination despite a worse calibration as compared with the ones described in the original studies that generated them. Nonetheless, such calibration is often improved when customizations are performed both at ICU and region or country level.

Summary
ICU scoring systems provide a valuable framework to characterize patients' severity of illness for the evaluation of ICU performance, for quality improvement initiatives and for benchmarking purposes. However, to ensure the best accuracy, constant updates as well as regional customizations are required.

Keywords
Acute Physiology and Chronic Health Evaluation, benchmarking, MPM0-III, Simplified Acute Physiology Score, scoring systems

INTRODUCTION

The Acute Physiology and Chronic Health Evaluation (APACHE), the Mortality Probability Model (MPM) and the Simplified Acute Physiology Score (SAPS) are the three most frequently used general severity-of-illness scores in adult ICU. The first generations of these scoring systems were introduced in critical care during the 1980s and were quickly incorporated into medical practice by intensivists [1–4]. They encompass clinical data regarding previous health status and the main acute diagnosis as well as physiologic and laboratory data by the ICU admission to estimate patients' outcomes, most invariably the vital status at hospital discharge. Although such instruments are of little assistance to the management of individual patients, they have been used by clinicians, researchers and administrators in the field of critical care to characterize patients in terms of severity of illness in clinical studies, for the evaluation of ICU performance, in quality improvement initiatives and for benchmarking purposes, among other potential uses.

Over the early 1990s, updated versions of the three scores were developed using data from a larger number of patients and employing more sophisticated statistical analyses [5–7]. Scoring system updates are often required, as it was demonstrated that the performances of these instruments suffer deterioration over time, as characterized by the worsening of discrimination (i.e., the capacity to discriminate survivors and nonsurvivors) and

[a]D'Or Institute for Research and Education and [b]Programa de Pós-Graduação em Oncologia, Instituto Nacional de Câncer, Rio de Janeiro, Brazil

Correspondence to Jorge I.F. Salluh, MD, PhD, D'Or Institute for Research and Education, Rua Diniz Cordeiro, 30 – 3 andar, Rio de Janeiro, CEP 22281-100, Brazil. Tel: +55 21 3883 6000; fax: +55 21 3883 6000; e-mail: jorgesalluh@gmail.com

Curr Opin Crit Care 2014, 20:557–565
DOI:10.1097/MCC.0000000000000135

KEY POINTS

- APACHE-IV, MPM0-III and SAPS 3 scores are useful to evaluate the outcomes and characterize disease severity of ICU patients.
- To ensure the best calibration, updates and regional customizations are required for ICU scoring systems.
- Studies demonstrate that such instruments are useful for the evaluation of ICU performance, in quality improvement initiatives and for benchmarking purposes.

calibration (i.e., the agreement between the observed and expected numbers of survivors and nonsurvivors across all of the strata of probabilities of death) [8,9]. Such deteriorations in terms of model performance (more importantly, of calibration) can be ascribed mainly to changes in case-mix and advances in both patients' management and science [10]. Over the last decade, the third or fourth generation of these scores, namely the APACHE IV, MPM0-III and SAPS 3, was developed using very large databases [9,11–13]. In the present article, we review the recent literature about these third or fourth-generation scoring systems. Readers can refer to an accompanying article in the present issue of the journal [14] and to an other three comprehensive reviews addressing aspects related to the development, validation and potential strengths and limitations of severity scoring systems in critical care [15–18].

THE ACUTE PHYSIOLOGY AND CHRONIC HEALTH EVALUATION IV, MPM0-III AND SIMPLIFIED ACUTE PHYSIOLOGY SCORE 3 SCORING SYSTEMS

There are several common characteristics among the APACHE IV, MPM0-III and SAPS 3 scores, partly because they were updated due to similar concerns. These scores were developed using prospectively collected data from a large number of patients and more sophisticated statistical analyses, and used information at the beginning of ICU stay to estimate the probability of hospital mortality. Nevertheless, they present several different characteristics. The development of APACHE IV and MPM0-III used data exclusively or predominantly [because four (3%) of the ICUs in the MPM0-III study were from Canada and Brazil] from the United States, respectively [9,13]. Both scores are proprietary tools owned by Cerner Corporation (Kansas City, Missouri, United States), although the company has made the scores and prediction algorithms publicly available. Conversely, the SAPS 3 score resulted from a multicenter study carried out in 35 countries worldwide, and this initiative was endorsed by the European Society of Intensive Care Medicine [11,12]. Table 1 summarizes the main characteristics of each scoring system.

The APACHE IV uses a large amount of data from the first day of ICU admission, including 116 specific acute diagnoses. SAPS 3 and MPM0-III use data exclusively obtained at the time of ICU admission (± 1 h) as their proponents have focused on the simplicity and feasibility of their routine use. Therefore, the abstraction burden of APACHE IV is substantially greater than that of SAPS II and MPM0-III scores [19], making these scores more suitable for ICUs where data collection is manual. Nonetheless, as systems and device interfaces are progressively being adopted in ICUs, automatic data extraction should minimize the data abstraction burden. Concerns regarding the labor-intensive task of calculating scores led investigators to create an automated ICU score, on the basis of the SAPS 3, exclusively using data available in the electronic medical record. In this study involving 67 889 ICU admissions at 21 hospitals between 2007 and 2011 the customized eSAPS 3 score demonstrated good discrimination [area under the receiver operator curve (AROC) = 0.82] and calibration (Hosmer-Lemeshow, $P = 0.57$) [20]. Moreover, by using data at ICU admission, MPM0-III and SAPS 3 estimates are less prone to influences related to in-ICU interventions. Interrater (interobserver) variability has been a source of concern in computing the scores, and investigators claim the APACHE scores are more prone to it [21]. Conversely, the interobserver variability seems to be adequate for SAPS scores [22]. Although there are no specific studies on this topic, we believe that the same must be true for MPM0-III. The main advantages and shortcomings for each model are shown in Table 2.

The evaluation of resource use is paramount for assessing quality in ICU, and the ICU length of stay (LOS) has been used as a proxy of resource use in ICU. The APACHE IV provides prediction equations to estimate the ICU LOS [9]. Although not present in its original scope, the SAPS 3 has also been used to examine variability in resource use between ICUs [1–4,23]. The investigators used a different approach to assess the standardized severity-adjusted resource use for each individual ICU. In this case, severity-adjusted resource use estimates the average amount of resources used per surviving patient in a specific ICU. More recently, investigators from the California Intensive Care Outcomes project developed a customized MPM0-III model to estimate ICU LOS [5–7,19]. However, it should be reinforced that, analogously to mortality prediction, ICU LOS estimates

Table 1. Main characteristics of the developmental studies for the Acute Physiology and Chronic Health Evaluation-IV, MPM0-III and Simplified Acute Physiology Score 3 scores

Scoring system	Patients and setting	Required variables	Time of data collection	Estimated parameters	Performance in the original validation set
APACHE-IV [9]	110 518 patients admitted to 104 ICUs from the United States between January 2002 and December 2003	Physiologic data ($n=17$), ICU admission diagnosis ($n=116$), chronic health variables ($n=6$), age, hospital location and LOS before admission, emergency surgery, thrombolytic therapy, mechanical ventilation	First ICU day	APACHE-IV score and predicted hospital mortality and ICU LOS	AROC: 0.880 H-LGOF C statistics: 16.8 ($P=0.08$)
MPM0-III [13]	124 885 patients admitted to 135 ICUs predominantly from the United States between October 2001 and March 2004*	Physiologic data ($n=3$), acute ($n=5$) and chronic ($n=3$) diagnoses, age, hospital location and LOS before admission, vasopressors use before ICU admission, type of admission, infection at ICU admission	At ICU admission ($\pm$ 1 h)	Predicted hospital mortality and ICU LOS**	AROC: 0.823 (95% CI, 0.818–0.828) H-LGOF statistics: 11.62 ($P=0.31$)
SAPS 3 [11,12]	19 577 patients admitted to 307 ICUs from 35 countries in five continents over a two-month period in 2002	Physiologic data ($n=10$), acute diagnosis and anatomical site of surgeries ($n=15$), chronic diagnoses ($n=6$), age, hospital location and LOS before admission, vasopressors use before ICU admission, type of admission, infection at ICU admission	At ICU admission ($\pm$ 1 h)	SAPS 3 score and respective predicted hospital mortality Customized equations for seven different geographic regions	AROC: 0.848 (95% CI, 0.841–0.854) H-LGOF C statistics: 14.29 ($P=0.16$); H-LGOF C statistics:10.56 ($P=0.39$)

APACHE, Acute Physiology and Chronic Health Evaluation; AROC, area under the receiver operator curve; CI, confidence interval; H-LGOF, Hosmer-Lemeshow goodness-of-fit; LOS, length of stay; MPM, Mortality Probability Model; SAPS, Simplified Acute Physiology Score.

*In the MPM0-III study, three ICUs were Canadian and one was Brazilian [13].

**Although there is no general standardized equation to estimate ICU LOS, MPM0-III was demonstrated to predict it by some investigators [19].

Table 2. Main advantages and disadvantages for the Acute Physiology and Chronic Health Evaluation-IV, MPM0-III and Simplified Acute Physiology Score 3 scores

Scoring system	Advantages	Disadvantages
APACHE-IV [9]	Coefficients regularly updated	Developmental sample restricted to one country
	Provides algorithms for LOS prediction	More complex data collection
	Specific algorithm to predict mortality in CABG surgery patients	High abstraction burden
	Less prone to be affected by the case-mix	Proprietary scoring system*
MPM0-III [13]	Lowest abstraction burden	Developmental sample mostly restricted to one country
	Less prone to interobserver variability	More susceptible to case-mix effects
	By using less physiologic data, may be preferred when laboratory resources are constrained	
SAPS 3 [11,12]	Low abstraction burden	Does not provide estimation for LOS
	Less prone to interobserver variability	Some regional equations were developed using relatively low sample size
	Developmental sample from 35 countries in five continents	
	Customized equations to predict hospital mortality according to seven different geographic regions	
	Potential use for international benchmarking	

APACHE, Acute Physiology and Chronic Health Evaluation; CABG, coronary artery by-pass graft; LOS, length of stay; MPM, Mortality Probability Model; SAPS, Simplified Acute Physiology Score.
*Cerner Corporation has recently made the score algorithms publicly available.

should not be used on an individual basis but as a measure to assist in the evaluation of ICU performance.

VALIDATION STUDIES IN GENERAL INTENSIVE CARE UNIT PATIENTS AND REGIONAL VALIDATIONS

Over the past decade, investigators evaluated the validation of these scores in different geographic regions and specific settings. Additionally, a comparison with the 'older generation' scores (e.g., SAPS II, APACHE II and III) was often performed. Several validation studies of the APACHE IV, SAPS 3 and MPM0-III scores were reported over the last years with conflicting results, as expected. As a consequence, some proposed not only their validation but also customizations aiming to improve their performance. These customized versions were made either at institutional or region or country level. The vast majority of customizations occur at first level (i.e., to compute a new logit in the regression equations) and refer to the SAPS 3 score. Of note, customized equations for seven different geographic regions worldwide were developed and made available in the original SAPS 3 report [8,9,11,12]. In this section of the article, we focus on validation studies in general ICU patients. Validation studies in specific patient subgroups and population are discussed in the next section.

As stated before in this article, the original calibration in the third or fourth-generation scores was good, as well as its discrimination, ranging from 0.82 to 0.88 [9–13]. One of the first validation studies was performed in a single ICU in Belgium where Ledoux *et al.* evaluated the performance of SAPS 3 as compared with APACHE II and SAPS II [9,11–13,24]. In this study, 851 consecutive patients were enrolled and the authors observed that AROC of the APACHE II model was significantly lower than for the SAPS II and SAPS 3 models. A good calibration was observed only for SAPS II and the SAPS 3 model customized for Central and Western Europe [14,24]. There are several studies that performed an external validation of the SAPS 3 and APACHE IV scores from different countries; however, most of these data are restricted to a small number of ICUs or small sample size and limited patient case-mix [15–18,24–32]. Auspiciously, data from multicenter studies involving a large number of patients are available. A study in 28 357 patients from 147 Italian ICUs demonstrated that SAPS 3 had a good discrimination but poor calibration, potentially limiting its use for benchmarking of ICUs in Italy [9,13,33]. Additionally, a subsequent multicenter Austrian study confirmed these findings [11,12,34]. In this study, the authors performed a regional customization of the score and concluded that region-specific or country-specific equations may be helpful to improve its use for benchmarking

purposes. Poole *et al.*, evaluating 2 661 patients from 103 Italian ICUs, also demonstrated that SAPS 3 overpredicted mortality even when compared with SAPS II [19,35]. Clearly, the role of SAPS 3 for benchmarking purposes has a patent geographic variation. For instance, in Brazil, the SAPS 3 has been recommended as the preferential severity of illness score by the Brazilian Association of Intensive Care (AMIB) since 2009. Data from approximately 200 000 patients, who were admitted during 2013 to 482 ICUs (1/3 of all adult ICU beds in Brazil) using the largest local ICU database for benchmark purposes, indicated a standardized mortality ratio of 1.03 [95% confidence interval (CI), 1.01 – 1.05] using the customized equation for South America and Caribbean countries of the SAPS 3 score [20,36]. Good discrimination and calibration were also shown for APACHE IV and MPM0-III in a large database of ICU patients in 21 North American hospitals [20,21]. APACHE IV and MPM0-III were validated in a multicenter study of 11 300 ICU patients from the United States and APACHE IV had better discrimination as compared with MPM0-III [19,22]. MPM0-III was also validated in another large database (of 55 459 patients) from 103 North American ICUs [9,37]. Brinkman *et al.* [38] performed an external validation of the APACHE IV and compared it with APACHE II and SAPS II in 62 737 patients from 59 Dutch ICUs. In a similar way to the studies performed with the SAPS 3 score, the authors demonstrated that, although the APACHE IV presented a good discrimination (AROC = 0.87), calibration was poor but significantly improved after customization. Studies comparing APACHE-IV, MPM0-III and SAPS 3 scores are summarized in Table 3.

Taken together, these studies are important as they reflect that, although the last generation of scoring system may consistently be applied in most ICUs, there is nonetheless room for improvement, customization or update in the current scores. Certainly, several aspects may help explain the divergence in discrimination and calibration from the original studies to those found in subsequent validations, namely differences in case-mix, process of care and resuscitation status (lead time bias), frequencies of do-not-resuscitate orders, source and type of data entry (administrative, software or manual) as well as ICU admission and discharge policies. Another aspect that is always to be considered is the potential deterioration of the system performance over time, indicating its need to be recalibrated.

VALIDATION STUDIES IN SPECIFIC SUBGROUPS OF PATIENTS

In the past years, several studies were performed evaluating the performance of APACHE IV, MPM0-III and SAPS 3 in specific subgroups of critically ill patients. In the sections below, we describe studies on cancer and solid-organ transplant patients as well as those requiring renal replacement therapy or extra-corporeal membrane oxygen (ECMO), patients with acute coronary syndromes, postcardiac surgery and those that had a cardiac arrest.

Table 3. Selected external studies comparing the Acute Physiology and Chronic Health Evaluation-IV, MPM0-III and Simplified Acute Physiology Score 3 scores in predicting hospital mortality

References	Patients (*n*)	Design and setting	Main findings
Keegan *et al.* [15]	2596	Retrospective; three ICUs at one hospital in the United States	Discrimination was better for APACHE IV (AROC = 0.868) than SAPS 3 (AROC = 0.801) and MPM0-III (AROC = 0.721). However, calibration was poor for all models.
Juneja *et al.* [39]	653	Retrospective; one medical ICU in India	Discrimination was excellent (AROC > 0.9) and calibration was appropriate for all models. Predicted mortality provided by the APACHE IV was closer to the observed mortality, whereas both SAPS 3 and MPM0-III overestimated mortality.
Nassar Jr. *et al.* [30]	5780	Retrospective; three ICUs in Brazil	Discrimination was very good for all models (APACHE IV, AROC = 0.883; SAPS 3, AROC = 0.855; MPM0-III, AROC = 0.840), but superior to APACHE IV. However, all models calibrated poorly and overestimated hospital mortality.
Kuzniewicz *et al.* [19]	11 300	Retrospective; ICUs from 35 hospitals in the United States	Discrimination was very good for all models; APACHE IV had the best discrimination (AROC = 0.892) compared with MPM0-III (AROC = 0.809) and SAPS II (AROC = 0.873). Calibration was good for MPM0-III, but not for APACHE IV. Abstraction time was three times longer for APACHE IV than for MPM0-III.

APACHE, Acute Physiology and Chronic Health Evaluation; AROC, area under the receiver operator curve; MPM, Mortality Probability Model; SAPS, Simplified Acute Physiology Score.

Patients with cancer and solid-organ transplant

The first external validation of the SAPS 3 score was performed by our group in a retrospective single-center study with 952 patients [40]. In that study, the customized equation for Caribbean and South America had the best performance and accurately predicted hospital mortality, even when scheduled surgical patients were excluded. Some years later, our group performed a new validation in a prospective study including patients admitted to 28 Brazilian ICUs and comparable results for the SAPS 3 were observed [41]. In this latter study, the MPM0-III score had a poor performance and tended to underestimate mortality in critically ill cancer patients.

The SAPS 3 was also evaluated along with the APACHE II in 501 patients who had undergone different solid-organ transplants and its performance was considered inadequate [42].

Patients requiring renal replacement therapy and extra-corporeal membrane oxygen

Three studies from Taiwan and two from Brazil evaluated the scores in patients requiring ECMO and/or renal replacement therapy [43–47]. Of note, the performance of APACHE IV and customized equations of SAPS 3 scores at ECMO or renal replacement therapy start was better than at ICU admission. These findings, however, were not present when using the MPM0-III score [45,48]. Moreover, all scoring systems performed poorly when estimated at ICU admission in this subgroup of patients.

Postcardiac arrest, coronary and cardiac surgical patients

Three validation studies were performed in cardiac patients. In a recent study by Doerr *et al.* [49], SAPS II and 3 scores had average discrimination and poor calibration. Two studies from Thailand and Brazil demonstrated that the SAPS 3 score is inaccurate in patients with acute coronary syndromes [50,51]. The APACHE IV score had both good performance and calibration, comparable to the more specific Global Registry of Acute Coronary Events (GRACE) score, despite a trend to overestimate mortality [51]. Two studies found a poor performance for the SAPS 3 score in postcardiac arrest patients [52,53].

MAIN RECOMMENDATIONS AND FUTURE DIRECTIONS

Considering the above-mentioned strengths and limitations of the scoring systems, one can conclude that, for both outcome prediction and benchmarking, the best option would be to use a score that was developed and validated recently in the country (or geographic region) where it will be employed (as is the case of the Intensive Care National Audit and Research Centre (ICNARC) case-mix program model [54]). Even in this case, periodic update of the score should be performed to reflect changes in medical care, as well as the changes in case-mix over time [15]. Although this provides a score with excellent local use, the downside to this approach relates to the lack of potential use for international comparisons and even for risk assessment in international clinical studies (where a common score would have to be used). An interesting approach that potentially mitigates this limitation is the use of a scoring system that is developed and validated using data from several international sites and that adjusts for the significant differences by using singular equations for each of the geographic regions (as in the case of SAPS 3).

At ICU or hospital level, when choosing a scoring system, other aspects that should be taken into account are as follows: feasibility (e.g., time to calculate the score, open versus copyright-protected scores), interobserver variability and the performance of the score in the case of specific populations. In our opinion, because of the fact that the present scores have heterogeneous performance in specific populations (Table 4), the choice of the scoring system should be based on its characteristics when applied to general ICU patients. In this case, specialized scoring systems (e.g., GRACE for acute coronary syndromes, European System for Cardiac Operative Risk Evaluation (EUROSCORE) 2 for cardiac surgery, among others) should be employed for specific populations when deemed necessary (e.g., when these patients represent a high volume of ICU admissions or there are ongoing disease-specific quality improvement programs). Also, at present, as reliance on scoring systems alone may not be sufficient to provide solid data on the performance of the ICU, collecting data on adherence to process of care measures has been increasingly recommended.

Finally, we believe that future versions of current scoring systems as well as new scoring systems to be developed should assess outcomes other than hospital mortality and integrate data on resource utilization, LOS, readmission and potentially long-term outcomes. Also, new scores should benefit from the huge amount of individual patient data that is now available in electronic medical records, ICU monitors and other medical devices. This should allow better individual profiling and perhaps hopefully improving individual patients' outcomes assessment.

Table 4. Selected external studies of the Acute Physiology and Chronic Health Evaluation-IV, MPM0-III and Simplified Acute Physiology Score 3 scores in predicting hospital mortality for specific subgroups of patients or special populations

References	Scores	Patients (*n*)	Design and setting	Main findings
Nassar Jr *et al.* [51]	SAPS 3, APACHE IV, GRACE Score	1065 patients with acute coronary syndromes	Retrospective; three ICUs in Brazil	Discrimination was very good for all evaluated scores (GRACE, AROC = 0.862; APACHE IV, AROC = 0.860; SAPS 3, AROC = 0.804). Calibration was appropriate for APACHE IV and GRACE but not for SAPS 3. All scores tended to overestimate mortality.
Costa-e-Silva *et al.* [48]	SAPS 3, APACHE IV, MPM0-III	366 with AKI	Six ICUs at one hospital in Brazil	Scores were assessed on nephrology consultation day. Discrimination was good, in general. Calibration was appropriate for all models, but not for MPM0-III. Mortality prediction by the SAPS 3 using the CSA customized equation was accurate, while the other scores underestimated mortality.
Wu *et al.* [44]	APACHE II, III and IV	102 patients supported by ECMO and acute dialysis	Retrospective; multicenter study in Taiwan	Discrimination was poor for all APACHE versions (APACHE IV had the highest AROC = 0.653). In this small study, as expected, calibration was appropriate for all models.
Lin *et al.* [47]	APACHE III and IV	78 patients supported with ECMO	Retrospective; one ICU in Taiwan	APACHE IV at ECMO initiation presented with excellent discrimination (AROC = 0.922) and good calibration.
Bisbal *et al.* [53]	SAPS 3, SAPS II, SOFA and Out-of Hospital Cardiac Arrest Score (OHCA)	124 postcardiac arrest patients admitted to the ICU	Retrospective; one center in France	SAPS 3 had poor discrimination (AROC = 0.62) and underestimated hospital mortality.
Doerr *et al.* [49]	SAPS II and 3	5207 postcardiac surgery patients	Retrospective; one center in Germany	On Day 1, discrimination was higher for SAPS II (AROC = 0.804) than for SAPS 3 (AROC = 0.757). Calibration was poor for both models.
Oliveira *et al.* [42]	APACHE II and SAPS 3	501 transplant patients	Retrospective; one center in Brazil	Calibration and discrimination were poor for the two evaluated scores. Discrimination was also poor when all types of transplants were evaluated separately.
Salciccioli *et al.* [52]	SAPS II and 3	274 postcardiac arrest patients	Retrospective; one center in the United States	Discrimination was not good for both scores (SAPS II, AROC = 0.70; SAPS 3 AROC = 0.66). Calibration was not assessed.
Khwannimit and Bhurayanontachai [50]	APACHE II, SAPS II and 3	2022 coronary care patients	Retrospective; one center in Thailand	Discrimination was excellent for all models (AROC > 0.9). However, all three models underestimated mortality and had poor calibration. The calibration of all scores was improved by first-level customization. All scores overestimated mortality.
Soares *et al.* [41]	SAPS II and 3, MPM0-III and CMM	717 patients with cancer	Prospective; 28 centers in Brazil	The SAPS 3 customized equation for CSA was accurate in predicting outcomes, with both good calibration and discrimination. SAPS II and 3 had the highest discrimination (AROC = 0.84), and MPM0-III the lowest (AROC = 0.71). Calibration was better using CMM and the customized equation of SAPS 3. The other general scores underestimated mortality while CMM tended to overestimate it.

(*Continued*)

Table 4 (*Continued*)

References	Scores	Patients (*n*)	Design and setting	Main findings
Maccariello *et al.* [45]	MPM0-III and SAPS 3	244 patients with AKI in need for dialysis	Prospective; 11 ICUs at three hospitals in Brazil	Scores were estimated at the start of dialysis. Discrimination was better for SAPS 3 (AROC: 0.82) and calibration was adequate when using the CSA customized equation. Discrimination (AROC: 0.73) and calibration were both worse using MPM0-III. SAPS 3 accurately predicted mortality, but the MPM0-III underestimated it.
Tsai *et al.* [46]	APACHE II, SAPS II and 3, SOFA and MODS	104 patients supported by ECMO and acute dialysis	Retrospective; multicenter study in Taiwan	Although the SAPS 3 score at dialysis start had the highest AROC (0.73), discrimination was poor for all other scores regardless of being estimated at ICU admission or at dialysis commencement. SAPS 3 over- and underestimated mortality in low- and high-risk patients, respectively.
Soares and Salluh [40]	SAPS II and 3	952 patients with cancer	Retrospective; one ICU in Brazil	Discrimination was very good for SAPS II and 3 (AROCs of 0.88 and 0.87, respectively). Calibration was appropriate for SAPS 3, particularly using the CSA customized equation, but not for SAPS II. The SAPS 3 CSA customized was accurate in predicting mortality, while the SAPS II underestimated it.

AKI, acute kidney injury; APACHE, Acute Physiology and Chronic Health Evaluation; AROC, area under the receiver operator curve; CMM, Cancer Mortality Model; CSA, Caribbean and South America; ECMO, extracorporeal membrane oxygen; MODS, Multiple Organ Dysfunction Score; MPM, Mortality Probability Model; SAPS, Simplified Acute Physiology Score; SOFA, Sequential Organ Dysfunction Score.

CONCLUSION

ICU scoring systems, such as APACHE IV, MPM0-III and SAPS 3, represent significant advances in comparison with the earlier generation of scores. The scoring systems have been thoroughly studied and validated and currently provide a valuable framework to characterize patients' severity of illness for the evaluation of ICU performance, for quality improvement initiatives and for benchmarking purposes. However, to ensure the best accuracy, constant updates as well as regional customizations are required.

Acknowledgements

Financial support: The authors are supported in part by grants from the National Council for Scientific and Technological Development (CNPq) and Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ).

Conflicts of interest

The authors are founders and equity shareholders of Epimed Solutions, which commercializes the Epimed Monitor, a cloud-based software for ICU management and benchmarking.

REFERENCES AND RECOMMENDED READING

Papers of particular interest, published within the annual period of review, have been highlighted as:

- ■ of special interest
- ■■ of outstanding interest

1. Knaus WA, Zimmerman JE, Wagner DP, *et al.* APACHE-acute physiology and chronic health evaluation: a physiologically based classification system. Crit Care Med 1981; 9:591–597.
2. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. Crit Care Med 1985; 13:818–829.
3. Le Gall JR, Loirat P, Alperovitch A, *et al.* A simplified acute physiology score for ICU patients. Crit Care Med 1984; 12:975–977.
4. Lemeshow S, Teres D, Avrunin JS, Pastides H. A comparison of methods to predict mortality of intensive care unit patients. Crit Care Med 1987; 15:715–722.
5. Le Gall JR, Lemeshow S, Saulnier F. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. JAMA 1993; 270:2957–2963.
6. 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.
7. Lemeshow S, Teres D, Klar J, *et al.* Mortality Probability Models (MPM II) based on an international cohort of intensive care unit patients. JAMA 1993; 270:2478–2486.
8. Capuzzo M, Moreno RP, Le Gall J-R. Outcome prediction in critical care: the Simplified Acute Physiology Score models. Curr Opin Crit Care 2008; 14:485–490.
9. 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; 34:1297–1310.
10. Moreno RP. Outcome prediction in intensive care: why we need to reinvent the wheel. Curr Opin Crit Care 2008; 14:483–484.
11. Moreno RP, Metnitz PGH, Almeida E, *et al.* 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.
12. Metnitz PGH, Moreno RP, Almeida E, *et al.* SAPS 3–from evaluation of the patient to evaluation of the intensive care unit. Part 1: objectives, methods and cohort description. Intensive Care Med 2005; 31:1336–1344.
13. 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.

14. Zimmerman JE, Kramer AA. A history of outcome prediction in the ICU. Curr Opin Crit Care 2014; 20:550–556.
15. Keegan MT, Gajic O, Afessa B. Severity of illness scoring systems in the intensive care unit. Crit Care Med 2011; 39:163–169.
16. Vincent J-L, Moreno R. Clinical review: scoring systems in the critically ill. Crit Care 2010; 14:207.
17. Breslow MJ, Badawi O. Severity scoring in the critically ill: part 2: maximizing value from outcome prediction scoring systems. Chest 2012; 141:518–527.
18. Breslow MJ, Badawi O. Severity scoring in the critically ill: part 1–interpretation and accuracy of outcome prediction scoring systems. Chest 2012; 141:245–252.
19. 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.
20. Liu V, Turk BJ, Ragins AI, *et al.* An electronic Simplified Acute Physiology Score-based risk adjustment score for critical illness in an integrated healthcare system. Crit Care Med 2013; 41:41–48.
21. Polderman KH, Jorna EM, Girbes AR. Inter-observer variability in APACHE II scoring: effect of strict guidelines and training. Intensive Care Med 2001; 27:1365–1369.
22. Strand K, Strand LI, Flaatten H. The interrater reliability of SAPS II and SAPS 3. Intensive Care Med 2010; 36:850–853.
23. Rothen HU, Stricker K, Einfalt J, *et al.* Variability in outcome and resource use in intensive care units. Intensive Care Med 2007; 33:1329–1336.
24. Ledoux D, Canivet J-L, Preiser J-C, *et al.* SAPS 3 admission score: an external validation in a general intensive care population. Intensive Care Med 2008; 34:1873–1877.
25. Capuzzo M, Scaramuzza A, Vaccarini B, *et al.* Validation of SAPS 3 Admission Score and comparison with SAPS II. Acta Anaesthesiol Scand 2009; 53:589–594.
26. Strand K, Soreide E, Aardal S, Flaatten H. A comparison of SAPS II and SAPS 3 in a Norwegian intensive care unit population. Acta Anaesthesiol Scand 2009; 53:595–600.
27. Lim SY, Ham CR, Park SY, *et al.* Validation of the Simplified Acute Physiology Score 3 scoring system in a Korean intensive care unit. Yonsei Med J 2011; 52:59.
28. Khwannimit B, Bhurayanontachai R. The performance and customization of SAPS 3 admission score in a Thai medical intensive care unit. Intensive Care Med 2010; 36:342–346.
29. Silva Junior JM, Malbouisson LMS, Nuevo HL, *et al.* Applicability of the simplified acute physiology score (SAPS 3) in Brazilian hospitals. Rev Bras Anestesiol 2010; 60:20–31.
30. Nassar APJ, Mocelin AO, Nunes ALB, *et al.* Caution when using prognostic models: a prospective comparison of 3 recent prognostic models. J Crit Care 2012; 27:423; e1–7. doi: 10.1016/j.jcrc.2011.08.016.
31. Mann SL, Marshall MR, Woodford BJ, *et al.* Predictive performance of Acute Physiological and Chronic Health Evaluation releases II to IV: a single New Zealand centre experience. Anaesth Intensive Care 2012; 40:479–489.
32. Mbongo C-L, Monedero P, Guillen-Grima F, *et al.* Performance of SAPS3, compared with APACHE II and SOFA, to predict hospital mortality in a general ICU in Southern Europe. Eur J Anaesthesiol 2009; 26:940–945.
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.
34. Metnitz B, Schaden E, Moreno R, *et al.* Austrian validation and customization of the SAPS 3 admission score. Intensive Care Med 2009; 35:616–622.
35. Poole D, Rossi C, Latronico N, *et al.* Comparison between SAPS II and SAPS 3 in predicting hospital mortality in a cohort of 103 Italian ICUs. Is new always better? Intensive Care Med 2012; 38:1280–1288.
36. Epimed Monitor System. http://epimedmonitor.com. [Accessed 29 April 2014]
37. Higgins TL, Kramer AA, Nathanson BH, *et al.* Prospective validation of the intensive care unit admission Mortality Probability Model (MPM0-III). Crit Care Med 2009; 37:1619–1623.
38. 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; .e11-8. doi: 10.1016/j.jcrc.2010.07.007.
39. Juneja D, Singh O, Nasa P, Dang R. Comparison of newer scoring systems with the conventional scoring systems in general intensive care population. Minerva Anestesiol 2012; 78:194–200.
40. Soares M, Salluh JIF. Validation of the SAPS 3 admission prognostic model in patients with cancer in need of intensive care. Intensive Care Med 2006; 32:1839–1844.
41. Soares M, Silva UVA, Teles JMM, *et al.* Validation of four prognostic scores in patients with cancer admitted to Brazilian intensive care units: results from a prospective multicenter study. Intensive Care Med 2010; 36:1188–1195.
42. Oliveira VM, Brauner JS, Rodrigues Filho E, *et al.* Is SAPS 3 better than APACHE II at predicting mortality in critically ill transplant patients? Clinics (Sao Paulo) 2013; 68:153–158.
43. Costa e Silva VT, de Castro I, Liano F, *et al.* Performance of the third-generation models of severity scoring systems (APACHE IV, SAPS 3 and MPM-III) in acute kidney injury critically ill patients. Nephrol Dial Transplant 2011; 26:3894–3901.
44. Wu V-C, Tsai H-B, Yeh Y-C, *et al.* Patients supported by extracorporeal membrane oxygenation and acute dialysis: acute physiology and chronic health evaluation score in predicting hospital mortality. Artif Organs 2010; 34:828–835.
45. Maccariello E, Valente C, Nogueira L, *et al.* SAPS 3 scores at the start of renal replacement therapy predict mortality in critically ill patients with acute kidney injury. Kidney Int 2010; 77:51–56.
46. Tsai C-W, Lin Y-F, Wu V-C, *et al.* SAPS 3 at dialysis commencement is predictive of hospital mortality in patients supported by extracorporeal membrane oxygenation and acute dialysis. Eur J Cardiothorac Surg 2008; 34:1158–1164.
47. Lin CY, Tsai FC, Tian YC, *et al.* Evaluation of outcome scoring systems for patients on extracorporeal membrane oxygenation. Ann Thorac Surg 2007; 84:1256–1262.
48. Costa e Silva VT, de Castro I, Liano F, *et al.* Performance of the third-generation models of severity scoring systems (APACHE IV, SAPS 3 and MPM-III) in acute kidney injury critically ill patients. Nephrol Dial Transplant 2011; 26:3894–3901.
49. Doerr F, Badreldin AMA, Can F, *et al.* SAPS 3 is not superior to SAPS 2 in cardiac surgery patients. Scand Cardiovasc J 2014; 48:111–119.
50. Khwannimit B, Bhurayanontachai R. A comparison of the performance of Simplified Acute Physiology Score 3 with old standard severity scores and customized scores in a mixed medical-coronary care unit. Minerva Anestesiol 2011; 77:305–312.
51. Nassar Junior AP, Mocelin AO, Andrade FM, *et al.* SAPS 3, APACHE IV or GRACE: which score to choose for acute coronary syndrome patients in intensive care units? Sao Paulo Med J 2013; 131:173–178.
52. Salciccioli JD, Cristia C, Chase M, *et al.* Performance of SAPS II and SAPS III scores in postcardiac arrest. Minerva Anestesiol 2012; 78:1341–1347.
53. Bisbal M, Jouve E, Papazian L, *et al.* Effectiveness of SAPS III to predict hospital mortality for postcardiac arrest patients. Resuscitation 2014; 85:939–944.
54. Harrison DA, Rowan KM. Outcome prediction in critical care: the ICNARC model. Curr Opin Crit Care 2008; 14:506–512.

Optimal data systems: the future of clinical predictions and decision support

Leo A. Celi[a,*], *Marie Csete*[b,*], *and David Stone*[c,*]

Purpose of review
The purpose of the review is to describe the evolving concept and role of data as it relates to clinical predictions and decision-making.

Recent findings
Critical care medicine is, as an especially data-rich specialty, becoming acutely cognizant not only of its historic deficits in data utilization but also of its enormous potential for capturing, mining, and leveraging such data into well-designed decision support modalities as well as the formulation of robust best practices.

Summary
Modern electronic medical records create an opportunity to design complete and functional data systems that can support clinical care to a degree never seen before. Such systems are often referred to as 'data-driven,' but a better term is 'optimal data systems' (ODS). Here we discuss basic features of an ODS and its benefits, including the potential to transform clinical prediction and decision support.

Keywords
clinical, data mining, decision support systems, electronic health records, information systems

INTRODUCTION: SYSTEMS OF DATA

The 'age of information' combined with ubiquitous electronic medical records (EMRs) means, in theory, that all data necessary for optimal diagnosis, treatment, and prognostication can be available to clinicians. The EMR interfaced to scientific information creates both opportunity and considerable challenges in acquisition and presentation of the clinically relevant data, in ways that best inform decision-making. Within a single patient EMR, myriad data types are captured, identified and categorized, filtered, summarized and then employed to construct a dynamic and revisable assessment and treatment plan. The amount of data generated by a single patient in a single hospital admission, particularly in the ICU, is enormous. Currently, the way vast data are captured and entered into medical records, leveraged, and fed back to clinicians is far from optimal. Despite application of computational tools to support decision-making in similarly data-rich complex systems outside medicine, application of computational tools to clinical data is in its infancy. Care must be taken to design such systems strategically, with sufficient modifiability to accommodate innovative advances as novel data elements and underlying decisional principles are added, changed and deleted from the canon. The organization of clinical data systems, then, requires a framework architecture on which data at all levels of resolution can be logically arranged. Highly functional complex systems (both engineered and evolved) share common design features that should be considered in the rational design of clinical data systems. Such meticulously designed systems will usher in a new era in clinical predictions: the interest will expand from predicting outcomes at the patient level either for prognostication or to inform decisions, to predicting information gain from diagnostic tests and response to various treatment options for individual patients.

Arguably, data from outside the EMR can and should inform clinical decision-making. For example, continuous local pollution levels play a

[a]Massachusetts Institute of Technology, Cambridge, Massachusetts, [b]Huntington Medical Research Institutes, Pasadena, California and [c]University of Virginia School of Medicine, Charlottesville, Virginia, USA

Correspondence to Leo A. Celi, Massachusetts Institute of Technology 77 Massachusetts Avenue, E25-505 Cambridge, MA 02139, USA. Tel: +1 617 253 7937; e-mail: lceli@mit.edu

*Dr Leo A. Celi, Dr Marie Csete, and Dr David Stone contributed equally to this manuscript.

Curr Opin Crit Care 2014, 20:573–580
DOI:10.1097/MCC.0000000000000137

KEY POINTS

- The use of data in clinical decision-making can be thought of as a clinical data system in which the responsible clinician functions as the controller.
- The current era in which EMRs are nearly universally implemented provides an opportunity for optimizing data system design to capture and leverage data in ways not available to individual practitioners in a traditional article-based environment.
- An example of such design is real-time incorporation of vast data sources into the course of clinical workflow and decision-making.
- The data optimized system has the potential to improve outcomes by a variety of means, such as providing useful and reliable predictions, supporting standardized approaches to clinical problems, and leveraging the data available in both population clinical databases and information resources.
- Such meticulously designed systems will usher in a new era in clinical predictions: the interest will expand from predicting outcomes at the patient level either for prognostication or to inform decisions, to predicting information gain from diagnostic tests and response to various treatment options for individual patients.

role in health, but it is impractical to feed these back to physicians because no one knows what to do with them. In the future, one could envision a huge amount of information used for clinical purposes – various 'omics' databases, large longitudinal epidemiologic studies, clinical trials, basic and preclinical research – all (automatically) interfaced with the EMR and exploited for minute-to-minute predictions and decision support. But such a goal remains distant for now, and physicians continue to use clinical data items in much the way they did during the unconnected, paper-based world of the 20th century.

Here it is critical to note that, though the amount of data (collected but) hidden from clinical records is problematic, more data do not necessarily yield better predictions, decisions, or outcomes. Data organization around design principles is the key. For example, a list of every component of an airplane does not automatically yield a robust flying machine without engineering principles and controls. Similarly, future clinical record systems must be engineered with standardization at the core, customizability at the edges, the agility to accommodate changes in healthcare environment, and a software architecture that is robust and current yet modifiable without undue difficulties. Thus, although an optimized data-based care system is an ideal goal, its benefits are limited by the data available to the system, but more importantly by how the data are organized. For this reason, we focus on some near-term approaches to restructuring clinical data, as a system, from content that is currently available but not optimally employed in the context of decision-making.

Clinical care is based on data acquisition and analysis, but is not yet 'data-driven' in the stricter sense of being objective, systematic, structured, and replicable with the same best outcomes. In fact, the data deluge of clinical practice (and the medical literature) has made it progressively more difficult to be aware of all applicable data. Unpredictable outcomes – specifically those relating to interventions providing no value added to the patient, or, worse, adverse consequences – are far too common and do not lend themselves easily to medicine as an applied data science. Here we describe the current state of clinical data in an attempt to clarify and enhance the concept of what has often been referred to as a data-driven care system to leverage computing power to cope with, manage, and properly analyze just the right patient data in the context of the population. The ultimate goal is to improve decision-making for physicians and patients by providing predictions and individualized recommendations to reliably optimize patient outcomes.

ARCHITECTURE OF A CLINICAL DATA SYSTEM

A system is an interconnected and interacting assembly of components (a.k.a. modules, parts) that can perform functions not possible with just the individual components. The rules (or protocols) that dictate the range of behaviors of a system are designed in engineered systems, and evolved in biological systems. A system accepts inputs and processes them into outputs. The details of a controlled system's sensing, computation, and actuation are dictated by the particular architecture of that system.

Clinical data tend to consist mainly of modular elements. (Note that modular elements can be descriptive or diagnostic in nature, as well as therapeutic or interventional.) Clinical data format, however, is (increasingly) highly varied (single nucleotide polymorphisms, transcriptomes of a tumor biopsy, functional imaging, raw vs. transformed EEG signals, results of a diagnostic nerve block) and therefore difficult to integrate without new collection and analytic tools. For inpatients, data are collated in a per-stay medical record along with varying degrees of accompanying interpretation. For outpatients, data are often dispersed, less

well organized, and often functionally unavailable. For clinical data generally, no framework architecture is used for organizing data in the context of a physiologic system (e.g., neurologic) or a medical condition (e.g., sepsis).

The advent of enterprise EMRs that incorporate outpatient and inpatient functions has begun to address the issue of integrating the patient's entire data history. Nonetheless, in every encounter with a patient, the clinician's data view axis is restricted to the prior and current data of an individual patient, as well as to the education, experience, efficiency, and memory of the clinician (or clinical team). To a large and unacceptable degree, clinicians 're-invent the wheel' with every patient encounter.

The clinician is the controller of a clinical data system, and the patient is – in engineering terms – the 'plant' (Fig. 1). Thinking of the clinician as controller highlights the need to structure the input data for optimal output (diagnosis, intervention, collection of more data, and prediction). For the most part, the controller is the cerebral 'wetware' of the clinician, but expert analysis will be increasingly supplemented by automated clinical decision support modalities (http://www-03.ibm.com/innovation/us/watson/watson_in_healthcare.shtml; accessed 24 April 2014). Further sensing of the patient state (the plant) in response to actuation is fed back to the clinician. In engineered control systems (such as a thermostat), controllers are designed to iteratively re-examine and re-apply solutions to the 'plant,' a design that is also used by the clinician with feedback from treatment response incorporated back to close the loop.

Understanding clinical data (not just clinical care in the larger sense) in terms of optimized controllers is a fundamentally important concept for clinical data utility, as (healthy) physiology is dependent on well-studied physiologic control systems. Given the gap between a well-controlled system and current clinical practice, we propose the term 'optimal data system' (ODS) to distinguish current data-driven approaches from those that are purposefully designed. We envision ODS as an enhanced type of data-driven system, which selectively employs appropriate data elements to support formulation of the best possible decisions, including outcome prediction. These data do not only include the patient's own historical and current data, but will eventually incorporate pertinent population data findings, as well as decision support resources such as guidelines, preferably formulated without undue industrial or financial influence [1]. The idealized ODS would continuously assess and catalogue the resultant outcomes of clinical decisions to determine what are the best data and decisions that can be recommended in the future.

The ideal data system would also be organized in modules representing particular organs or disease states with these modules nested in the global data set reflecting system pathophysiology. Such organization and presentation are now in the hands of software developers, both a challenge and an opportunity. The ideal data system organization will

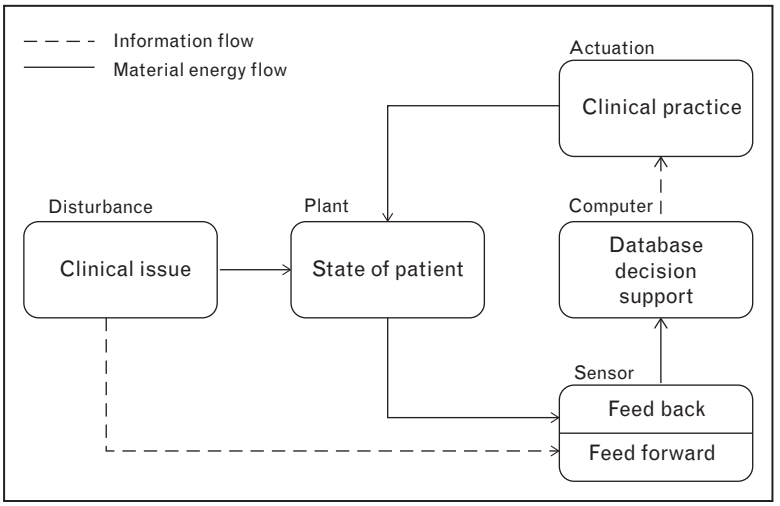

FIGURE 1. Control loop depicting a data-driven care system. A clinical issue such as an infection or vascular occlusion affects the state of the patient. Subsequently, the system sensor detects this change and submits the relevant data to the computer for storage and analysis. This results in actuation (or not) of a clinical practice intervention that further affects the state of the patient, which feeds back into the system for further analysis. Feed forward control involves the transmission of disturbances directly to the sensor without first affecting the state of the patient. The detection of a risk factor for venous thromboembolism that triggers prophylaxis in a protocol-based manner represents a clinical example of feed-forward control.

require intense feedback between these computer scientists and clinicians for the next generation of health information systems.

CONSTRAINTS AND TRADEOFFS IN THE UTILIZATION OF CLINICAL DATA

Medical care systems provide caregivers with various levels of opportunity to identify and acquire the data perceived as necessary at any given time for a given patient, but data identification is highly constrained. First, the data must be conceived and recognized as such, that is, identified as a clinical data element. Until the element has been established to be relevant to clinical care, the element will remain in the area of the Venn diagram that lies outside the clinical data area (Fig. 2). This may seem trivial, but once-essential data can become obsolete and completely new and unexpected forms of data become essential (e.g., troponins for myocardial infarction diagnosis today versus ~1978). Data may also simply be unknown to the user because of educational, experiential, or communication issues. Second, it must be recognized as valuable, that is, worthy of the cost of the acquisition and storage. Ideally, this value is established by studies that examine information gain of this particular element – does it lead to a better understanding of the disease process on top of what is already known and/or does it inform decision regarding a possible intervention that will alter patient outcome? Third, the data must be obtainable. The data may not be technically available because of a lack of equipment or because science has not yet established a method of examining the real-time function of a given gene or signal transduction pathway. Fourth, the data must be presented and formatted to the user in a timely manner (based on clinical acuity) and stored for future clinical and research utilization; these functions are facilitated and supported by EMRs. Much information is simply lost because it is not archived with the patient record (e.g., waveform signals, hemodialysis parameters) or functionally inaccessible in mounds of paper or microfilm.

Currently, the predominance of free text entry in physician notes makes the reliable cataloguing of data for future analysis for prediction and decision support rather difficult, but at least theoretically possible via tools such as natural language processing. The seamless integration of structured data capture in EMR workflow is still in early development. Medical research continues to identify novel and previously unrecognized elements from the

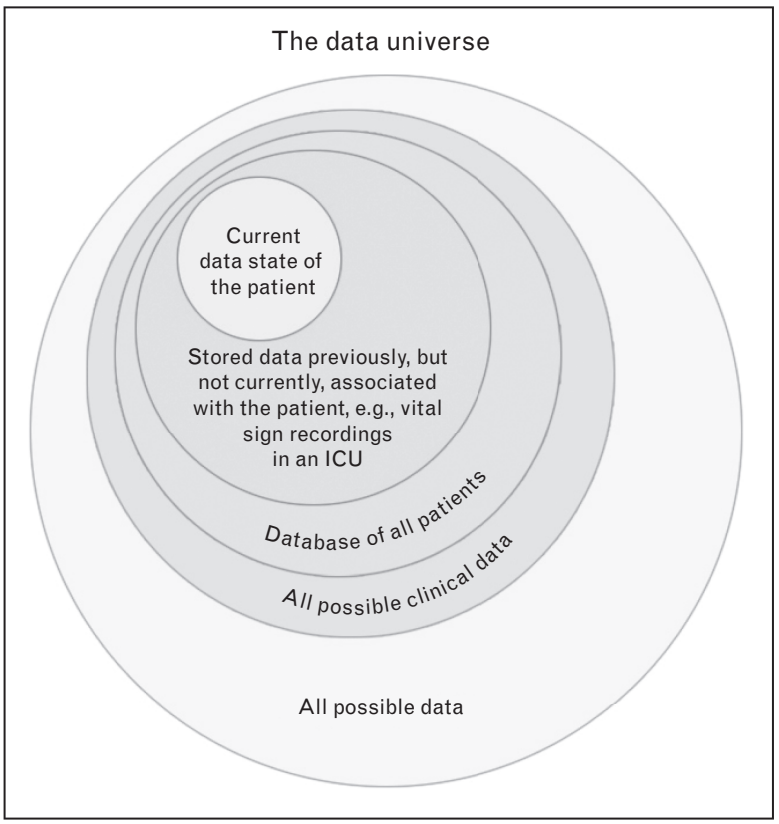

FIGURE 2. The Data Universe (not drawn to scale). Data move from the realm of 'all possible data' to that of 'all possible clinical data' as they are identified as having clinical value. Figure courtesy of Kai-ou Tang.

general universe of data, which become relevant to clinical care. Entirely new types of data may be developed in this fashion, for example, genetic testing for disease risk.

A need for 'all the data, all the time' can be wasteful, costly, confusing, and time-consuming. Clinicians who require an MRI for the evaluation of all back pain or headache cases will struggle to operate in environments where these are not available (or allowed), highlighting the important issues of cost, value, and risk of data. If one can function safely without such sophisticated diagnostic modalities in most circumstances, what is the appropriate threshold for using such modalities? At the other end of the spectrum, patients may suffer consequences from clinicians not performing tests. Clearly, we have not learned to capture 'just enough data,' which should be the goal of data systems design and evaluation.

Data may also be erroneous for a variety of reasons such as mis-entry, machine or human errors, unduly subjective circumstances, and limitations of medical device precision. Finally, data may go missing either because it was never entered or was lost in some quantum mechanical event occurring in a vast database over long time periods. Clearly, we, as all too human clinicians, need some help in identifying and utilizing data optimally.

CREATING AN OPTIMAL DATA SYSTEM

Acquisition of the necessary data elements as well as their subsequent assembly represents essential protocols of a data-driven system. The clinical puzzle is simply not as perfect as jigsaw pieces out of the box. Instead, pieces are missing and misshapen, and there may be strange extra pieces (Fig. 3). Even the final puzzle product can be a moving target. However, the formulation of some kind of mental construct built on data pieces is a useful model for the next steps of assessment, intervention, and reassessment. The ODS must be designed to support and facilitate an increasingly complex, rushed, and demanding clinical work environment.

We propose an enhancement to the current process of data incorporation into the decision

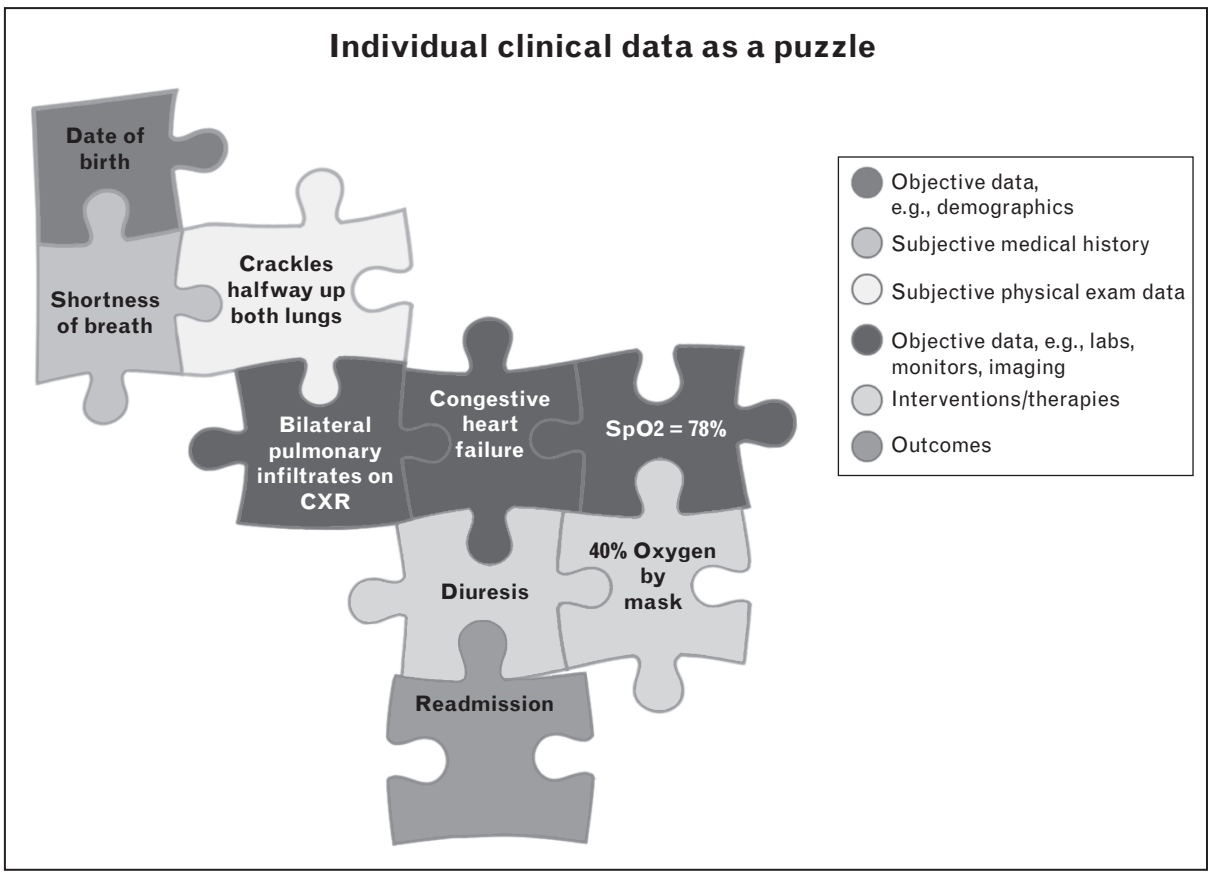

FIGURE 3. Individual clinical data as a puzzle. The puzzle changes as data are added/changed/removed, but the sequence of changes can be recapitulated by virtue of date/time stamping. Decision support by population database or practice guidelines could present options for new pieces, assembly suggestions, or deletion of pieces. CXR, chest X-ray. Figure courtesy of Kai-ou Tang.

and care process. As the nexus of clinical decisions is the medical note, the EMR is the logical platform in this development. An ideal ODS would include the following:

(1) Automatic collection and display of newly available data (i.e., data not yet entered in an EMR) required to complete the clinical picture. These could include patient-entered data; data sent by prehospital personnel real-time; and data from wearable sensors.
(2) Capture and integration of the newly available and historical data along with real-time physician entries (notes) to progressively characterize the clinical state and query both population database and clinical decision support modalities. These are represented by our dynamic clinical data mining (DCDM) concept and the IBM Watson type of functionality, respectively (http://www-03.ibm.com/innovation/us/watson/watson_in_healthcare.shtml; accessed 24 April 2014) [2▪▪]. This complex feature requires de-identified data sharing on a universal basis [2▪▪].
(3) An innovative additional feature would be required to integrate the DCDM and Watson functions and deliver the following: diagnostic, therapeutic, prognostic as well as further documentation suggestions would be incrementally displayed on the basis of the combination of the analyzed data provided by these multiple sources. These might include suggestions to supplement required missing data with additional testing; clarification of free text entries for purposes of standard coding; identification of suggestive but otherwise difficult-to-identify patterns and constellations of data; automatic highlighting of diagnoses, treatments, results and combinations of results that are incongruous or inconsistent; and providing population-based but individualized suggestions for ongoing care decisions and next steps. This is the stage where the software–wetware integration process is continuously enhanced by leveraging information outside the purview of today's clinical EMR (or paper chart) user.
(4) Machine learning would be employed to continuously improve the quality of the information presented to the user as the system 'learns' how clinicians employ the system in heterogeneous ways.
(5) Users are allowed to customize their own version of the application to the extent that standardization of data is not violated. In other words, the application design should be 'customized at the edges but standardized at the core', enabling users to have considerable but reasonable control over their interactions with the system [3▪]. Customization should not be permitted to the extent that it is difficult or near impossible for software engineers to investigate reported system errors and unanticipated events.
(6) Saved system data would then be provided to both the local and the population databases for ongoing analysis for real-time care and the objective formulation of clinical support modalities, including practice guidelines and research.
(7) Reports would be generated regarding user decisions in terms of consistency with best practices as suggested by the system.
(8) The system should be modifiable so that it can incorporate new and innovative modalities for clinical prediction and decision support.
(9) The system should be modifiable so that important new information can be brought to the 'head of the line' under certain urgent circumstances such as drug recalls, epidemics, disasters, and acts of terrorism.
(10) The system should be fully tested in prototype by expert users in parallel with the current care system before allowing it to be used in daily practice by regular clinicians. This testing will probe usability as well as detect the kinds of system errors that can only be exposed with use in a real clinical context.

Experienced clinicians make decisions with minimal or 'just enough' data – they realize that there are costs to obtaining unnecessary data. These costs include not only the obvious human, financial, and clinical risks of further testing, but also the inevitable distractions of information overload. The ODS also introduces the opportunity for either systematic review or random auditing of clinical decisions. These audits would review system as well as individual human performance. Such analysis is already starting as organizations incorporate tools that identify clinicians who obtain insufficient, excessive, or wrong data, and who make decisions identified as suboptimal under the care circumstances. Systems approaches to teaching medicine are clearly needed to prepare clinicians for optimal use of data systems.

WILL OPTIMAL DATA SYSTEMS IMPROVE OUTCOMES?

First, no changes in care based on current data or processes can transcend the therapeutic limitations

of current practice: creating an optimally data-driven care system is a necessary starting point in re-engineering medicine for the digital age, but it does not represent a clinical panacea. The limitations in our actuational capabilities put a firm glass ceiling on the outcome improvements that can be achieved and measured without the implementation of truly innovative treatment advances. However, better use of data does provide the promise of contributing to future advances in this regard [4], and, more importantly, cost-effective use of tests and treatments in the near term. As intensivists, we recognize that critical care medicine is a particularly data-rich area of medicine, but has not heretofore captured or utilized these data to a significant extent [5▪].

The intensive use of data should allow us to recognize patterns in the administration of care that may contribute to otherwise undetectable positive or negative impacts on outcomes. For example, if clinicians had real-time access to prior outcomes in comparable patients, they could adjust their care plans on the basis of previously successful approaches in large populations [2▪▪]. Clinicians could also adjust their practice on the basis of observations of negative effects that can only be detected by the study of large populations [6▪]. These effects may be subtle or only occur under circumstances of specific combinations of clinical context and interventions, and therefore will not generally be noted in the course of normal practice or even chart reviews.

More carefully designed data presentation might speed up the process in which clinicians review data. This might simply mean better and smarter graphical displays [7]. For example, a better-engineered presentation of those data elements that the clinician needs to know and can actually act upon could safely eliminate the need to review all the data entries, all the time (Fig. 4).

Data can provide the basis for more robust and standardized care decisions, especially in frequently encountered situations such as acute hypotension in the ICU [8▪]. In addition, we may be able to use diagnostic testing in a more selective and cost-effective manner [9]. Workflow should be better supported – for example, where checklists are employed, available data could populate the checklist to some extent. Clinicians could be notified of the presence of uncompleted checklist items in a

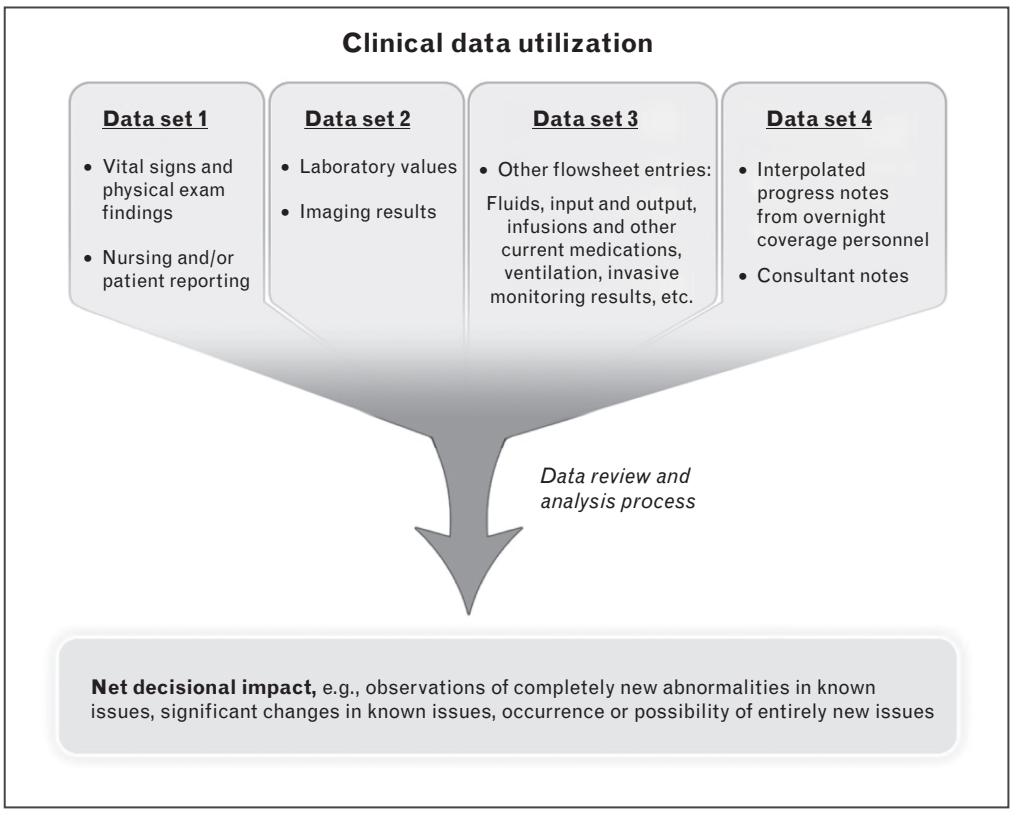

FIGURE 4. Clinical data utilization. The clinician may analyze dozens or hundreds of individual data items in the course of workflow, but only net a few significant data items that influence the course of decision making. The detection of zero change also influences the analysis. The issue raised here is how this iterative, detail-oriented process can be accelerated and supported by technology.

manner that should improve the accuracy and diminish the tedium of the task [10,11]. Any additional provision of time provided by a carefully designed data-driven system should itself provide an advantage as clinicians recognize time as a critical limiting factor in point of care practice [12,13].

Careful and directed use of data may allow us to discharge patients from the ICU more safely and efficiently [14]. Data can similarly be employed to identify patients with extremely poor prognoses who are receiving inevitably futile care [15]. The biggest impact of the data re-engineering is a more standardized decision-making based on predicted outcomes and retrospective comparative effectiveness analysis, avoidance of unnecessary testing, and unloading of provider cognitive workload to free up time that can be better spent on tasks that add value.

CONCLUSION

There is always a tension between practicing optimally on the basis of current knowledge and advancing the state of the art of patient care, which requires insights and interventions not yet in the canon. This tension is the result of an unnecessary gap between research and practice; clinicians currently execute this translational process without adequate data support. Clinicians also occasionally face decisions that must be made on an individual, experiential basis, as opposed to a more standardized approach, especially when patterns have no apparent precedent in that clinician's knowledge and experience. To complicate matters, new varieties and forms of data are incrementally added to clinical databases, as trials of new tests and therapies are known. The challenge to software designers and clinicians is incorporating the beneficial elements of these advances into an established information system firmly based on the integration of previously available individual and population data. Such advances will require algorithmic adjustment of the information presented to the user so that the impact of important discoveries is accelerated into a revisable and dynamically data-driven system of clinical practice.

Acknowledgements

We would like to acknowledge Ms Kai-ou Tang, who provided the figures presented in this paper.

Conflicts of interest

There are no conflicts of interest.

REFERENCES AND RECOMMENDED READING

Papers of particular interest, published within the annual period of review, have been highlighted as:

- ■ of special interest
- ■■ of outstanding interest

1. Steinbrook R. Improving clinical practice guidelines. JAMA Intern Med 2014; 174:181.
2. ■■ Celi LA, Zimolzak A, Stone DJ. Dynamic clinical data mining: search engine-based decision support. J Med Internet Res 2014; 2:e13.

This article describes the potential combination of EMRs with big data and search engines to provide real-time decision support for clinical issues not well addressed by available clinical trial results.

3. ■ Csete M, Doyle J. Bow ties, metabolism and disease. Trends Biotech 2004; 22:446–450.

This work provides further engineering background for those who wish to pursue this issue in more detail.

4. Denny JC, Bastarache L, Ritchie MD, *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nat Biotech 2013; 31:1102–1110.
5. ■ Celi LA, Mark RJ, Stone DJ, *et al.* Big data in the intensive care unit: closing the clinical data loop. Amer J Resp Crit Care Med 2013; 187:1157–1160.

This publication provides a background summary of the data issues relevant to critical care medicine.

6. ■ Ghassemi M, Marshall J, Singh N, *et al.* Leveraging an ICU database: increased ICU mortality noted with SSRI use. Chest 2013; 145:745–752.

This provides an example of the kind of observational study to supplement and complement randomized controlled trials that can be performed using clinical databases.

7. Tufte ER. The visual display of quantitative information. 2nd ed. Cheshire, CT: Graphics Press; 2001; 107–121.
8. ■ Mayaud L, Lai PS, Clifford GD, *et al.* Dynamic data during hypotensive episode improves mortality predictions among patients with sepsis and hypotension. Crit Care Med 2013; 41:954–962.

This is an example of the kinds of decision support tools that can be developed with more complete use of clinical data in the ICU.

9. Cismondi F, Celi LA, Fialho A, *et al.* Reducing ICU blood draws with artificial intelligence. Int J Med Inform 2013; 82:345–358.
10. Vincent JL. Give your patient a fast hug (at least) once a day. Crit Care Med 2005; 33:1225–1230.
11. Pronovost P, Barenholtz S, Dorman T, *et al.* Improving communication in the ICU using daily goals. J Crit Care 2003; 18:71–75.
12. Cook DA, Sorensen KJ, Wilkinson JM, *et al.* Barriers and decisions when answering clinical questions at the point of care: a grounded theory study. JAMA 2013; 173:1962–1969.
13. Westphal M. Get to the point in intensive care medicine – the sooner the better? Crit Care 2013; 17 (Suppl 1):S8.
14. Badawi O, Breslow MJ. Readmissions and death after ICU discharge: development and validation of two predictive models. PLoS ONE 2012; 7:e48758.
15. Huynh TN, Kleerup EC, Wiley JF, *et al.* The frequency and cost of treatment perceived to be futile in critical care. JAMA 2013; 173:1887–1894.