Critical Care

Evaluation of SOFA-based models for predicting mortality in the ICU: a systematic review

Critical Care 2008, **12**:R161 doi:10.1186/cc7160

Lilian Minne (lilian_minne@hotmail.com)
Ameen Abu-Hanna (a.abu-hanna@amc.uva.nl)
Evert de Jonge (e.dejonge@amc.uva.nl)

ISSN 1364-8535

Article type Research

Submission date 29 October 2008

Acceptance date 17 December 2008

Publication date 17 December 2008

Article URL http://ccforum.com/content/12/6/R161

Evaluation of SOFA-based models for predicting mortality in the ICU: A systematic review

Lilian Minne[1], Ameen Abu-Hanna*[1] and Evert de Jonge[2]

[1]Department of Medical Informatics, Academic Medical Center, Meibergdreef 9, 1105 AZ, Amsterdam, The Netherlands
[2]Intensive Care Department, Academic Medical Center, Meibergdreef 9, 1105 AZ, Amsterdam, The Netherlands

Email: Ameen Abu-Hanna*- a.abu-hanna@amc.uva.nl;

*Corresponding author

Abstract

Introduction : To systematically review studies that evaluate the performance of Sequential Organ Failure Assessment (SOFA)-based models for predicting mortality in ICU patients.

Methods : Medline, EMBASE and other databases were searched for English-language articles, whose major objective was to evaluate the prognostic performance of SOFA-based models in predicting mortality in surgical and/or medical ICU admissions. The quality of each study was assessed based on a quality framework for prognostic models.

Results : Eighteen articles met all inclusion criteria. The studies differed widely in the SOFA derivatives used and in their methods of evaluation. Ten studies reported on developing a probabilistic prognostic model, only five of which used an independent validation data set. The other studies used the SOFA-based score directly to discriminate between survivors and non-survivors without fitting a probabilistic model. In five out of six studies, admission-based models (APACHE II/III) were reported to have a slightly better discrimination ability than SOFA-based models at admission (the receiver operating characteristic curve (AUC) of SOFA-based models ranged between 0.61 and 0.88), and in one study a SOFA model had higher AUC than the SAPS II model. Four of these studies used the Hosmer-Lemeshow tests for calibration, none of which reported lack of fit for the SOFA models. Models based on sequential SOFA scores were described in eleven studies including maximum SOFA scores and maximum sum of individual components of the SOFA score (AUC range: 0.69 to 0.92) and delta SOFA (AUC range: 0.51 to 0.83). Studies comparing SOFA to other organ failure scores did not

consistently show superiority of one scoring system to another. Four studies combined SOFA-based derivatives with admission severity of illness scores, and they all reported on improved predictions for the combination. Quality of studies ranged from 11.5 to 19.5 points on a 20 point scale.

Conclusions : Models based on SOFA scores at admission had only slightly worse performance than APACHE II/III and were competitive with SAPS II models in predicting mortality in general medical and/or surgical ICU patients. Models with sequential SOFA scores seem to have a comparable performance to other organ failure scores. The combination of sequential SOFA derivatives with APACHE II/III and SAPS II models clearly improved prognostic performance of either model alone. Due to the heterogeneity of the studies, it is impossible to draw general conclusions on the optimal mathematical model and optimal derivatives of SOFA scores. Future studies should use a standard evaluation methodology with a standard set of outcome measures covering discrimination, calibration, and accuracy.

Introduction

The development of the Sepsis-related Organ Failure Assessment (SOFA) score was an attempt to objectively and quantitatively describe the degree of organ dysfunction over time and to evaluate morbidity in intensive care unit (ICU) septic patients [1]. Later, when it was realized that it could be applied equally well in non-septic patients, the acronym 'SOFA' was taken to refer to "Sequential Organ Failure Assessment (SOFA)" [2]. The SOFA scoring scheme daily assigns 1-4 points to each of the following six organ systems depending on the level of dysfunction: respiratory, circulatory, renal, haematology, hepatic and central nervous system. Since its introduction, the SOFA score has also been used for predicting mortality although it was not developed for this purpose.

The aim of this paper was to systematically review, identify research themes, and assess studies evaluating the prognostic performance of SOFA-based models (including probabilistic models and simple scores) for predicting mortality in medical and/or surgical adult Intensive Care Unit (ICU) patients.

Materials and methods

Search strategy

Two reviewers independently screened the titles and abstracts of articles obtained by the following search procedure. Scopus (Jan 1966 to February 2008), was searched for research articles and reviews using the

following query: *(critical OR intensive) AND (mortality OR survival) AND (sofa OR "sepsis-related organ failure" OR "sepsis related organ failure" OR "sequential organ failure"))* in title, abstract and keywords. Scopus comprises, among others, clinical databases like Medline and Embase. Only English journal articles were considered.

In addition, the references of all included articles as well as articles citing them were screened, and authors were approached about follow-up studies in progress. Follow-up studies were only included when they were already accepted for publication.

Inclusion criteria

The following inclusion criteria were applied: 1. the study aimed to evaluate a SOFA-based model (probabilistic or as a score), 2. it assessed the statistical performance of the model in terms of accuracy and/or discrimination and/or calibration (studies reporting only on odds ratios and/or standardized mortality ratios were excluded), 3. the predicted outcome of the study was mortality or survival of the patient, and 4. the patient sample was not restricted to a specific diagnosis (e.g. diabetes) but taken from the surgical and/or medical adult ICU population. Two reviewers conducted the search and differences were resolved by consensus after including a third reviewer.

Quality assessment

The quality of the included studies was assessed based on an adaptation of a quality assessment framework for systematic reviews of prognostic studies [3] [see Additional data file 1]. This framework includes the following 6 areas of potential study biases: 1. study participation, 2. study attrition, 3. measurement of prognostic factors, 4. measurement of and controlling for confounding variables, 5. measurement of outcomes, and 6. analysis approach. Two reviewers conducted the quality assessment independently from each other and discrepancies were resolved by involving the third reviewer.

Missing data

Authors were contacted by email to complete missing data that were required for characterizing the studies. When the authors did not reply or their answer was still unclear, empty fields were marked with 'Not Reported (NR)'.

Prognostic performance measures

For each included study we describe the reported discrimination of the model (or score) and if available the reported calibration and accuracy. Discrimination, usually measured in terms of the Area Under the Receiver Operating Characteristic Curve (AUC), refers to a model's ability to assign a higher probability to non-survivors than to survivors. The AUC, however, gives no indication of how close the predicted probabilities are to the true ones (estimated by the observed proportion of death). Calibration refers to this agreement between predicted and true probabilities and is most often measured by the Hosmer-Lemeshow H or C goodness-of-fit statistics (these are based on the χ^2 test). These statistics suggest good fit when the associated p-values are greater than 0.05, but they are strongly influenced by sample size. Accuracy is a measure of the average distance (residual) between the observed outcome and its predicted probability for each individual patient. A popular accuracy measure is the Brier score which is the mean squared residuals. The Brier score is sensitive to both discrimination as well as calibration of the predicted probabilities.

Results

Search results

Of 200 studies initially identified, 18 met the inclusion criteria and were included in this study (Figure 1). Inter-observer agreement measured by Kappa was 0.94.

By scanning the reference lists of included articles and those citing them, 7 additional articles were rendered potentially relevant. Nevertheless, assessment of their abstracts demonstrated that they did not match our inclusion criteria (6 studies did not provide data on discrimination, calibration or accuracy, and one study did not use SOFA to predict mortality).

[Here Figure 1]

Study characteristics

Table 1 shows the characteristics of the included studies. The studies evaluated different types of SOFA derivatives (e.g. mean, maximum (max)) and compared them to different models and covariates. Six studies combined SOFA with other models or covariates [4–9].

17 studies (94%) measured the AUC [4–7, 9–21], 4 studies (22%) measured the Brier score [4, 8, 9, 11], and 6 studies (33%) calculated Hosmer-Lemeshow (HL) statistics [4, 5, 7, 11, 14, 15] (two studies used the C-statistic [4, 11], one used the H-statistic [5], one used both [7], and the rest [14, 15] did not define which of the two statistics were used).

Studies were not always clear about the kind of model used to evaluate SOFA. Only ten studies (56%) reported the use of a logistic regression model [4–9, 14, 15, 20, 21]. The models in these studies were fitted on local developmental data sets. Five of these ten studies validated the model on an independent test set [4, 5, 8, 9, 15], and five studies did not report how the model was validated [6, 7, 14, 20, 21].
Hospital mortality was the outcome in 10 studies [4, 6, 8, 9, 11, 12, 14, 15, 17, 20], ICU mortality in 8 studies [5, 7, 10, 13, 14, 18, 19, 21], and in one study mortality was undefined [16]. One study evaluated both ICU and hospital mortality [14].

Missing data

Study characteristics that were most often missing were: type of patient population (surgical/medical/mix), type of model (e.g. logistic regression), and whether the model was validated on the developmental or independent validation set. Mailing the authors confirmed in the type of ICU outcome (hospital or ICU mortality) used in one study.

Study quality

We used 4 of the 6 main quality aspects in the framework of Hayden *et al.* [3] leaving 'study attrition' (such as loss to follow-up) and 'confounding measurement and account' out. The former is irrelevant in our analysis and the latter falls outside the scope of this review. The maximum quality score is 20. The results of the quality assessment of the included studies are shown in Table 2.

Study results

The cohort size ranged from 303 to 6,409 patients. Mean age was 53 to 62 years in complete cohorts, median age was 66 years in one study [15]. The percentage of males was 52% to 71%. Hospital mortality ranged from 11% to 45% and ICU mortality from 6.3% to 37%.
Studies were heterogeneous in the way they used SOFA. The major themes identified in the evaluation studies are investigating the performance of: Single SOFA scores at admission or at a fixed time after admission; Sequential measurements of SOFA (e.g. mean SOFA score); Individual components of SOFA (e.g. cardiovascular component); Combination of SOFA with other covariates; and Temporal models using patterns discovered in the SOFA scores.

Performance of single SOFA scores at a fixed time on and after admission

Eleven studies (61%) evaluated the SOFA score on admission (Table 3) [10–17, 19–21]. In seven studies, SOFA on admission was calculated using the most abnormal values from the first 24 hours after admission [10, 12, 14, 16, 17, 19, 20]. Discrimination, measured by the AUC, ranged between 0.61 and 0.88. P-values of HL-statistics ranged from 0.17 to 0.8. Four studies (22%) evaluated SOFA on other days than the day of admission [15–17, 19]. In these studies AUCs ranged between 0.727 and 0.897 and p-values of HL-statistics ranged between 0.09 and 0.27 for days 2-7 after admission and at the day of ICU discharge. Six studies (33%) compared admission SOFA with traditional admission-based models [11–13, 16, 17, 20]. The comparison is more meaningful in the first four studies [11, 12, 17, 20] which, in line with the admission-based models, were developed to predict hospital mortality. Two studies of them reported that the Acute Physiology And Chronic Health Condition (APACHE) II score had better or slightly better discrimination than admission SOFA [11–13]. Furthermore, one study found better calibration for the APACHE II score [11]. This same study also found that the Acute Physiology Score (defined as the APACHE II score without age and chronic health conditions) had comparable discriminative ability to admission SOFA, and better calibration. One study reported comparable discrimination (AUC = 0.776 and 0.825 for SOFA and APACHE III, respectively), and comparable calibration for SOFA and APACHE III on admission [17]. Finally, one study reported that admission SOFA had a higher AUC (0.82) than SAPS II (0.77) [20]. In the other two studies that compared admission SOFA with traditional admission-based models, the outcome was either ICU mortality [13] or undefined [16]. In these two studies the APACHE II score was reported to have slightly better discrimination than, but in essence comparable to, admission SOFA (0.62 vs. 0.61 in [13] and 0.88 vs. 0.872 in [16].

Five studies (28%) compared SOFA with other organ failure scores [10, 14–17]. Generally, no clear differences were found in calibration or discrimination (Table 3).

Performance of sequential measurements of SOFA

Eleven studies (61%) evaluated sequential measurements of SOFA [7, 11, 14–21]. The derivatives evaluated were: Max SOFA (4 studies), Total Max SOFA (7 studies), Delta SOFA (7 studies), Mean SOFA (2 studies), Total SOFA (1 study) and modified SOFA (2 studies) (Table 4).

Total Max SOFA was always defined as the sum of the highest scores per individual organ system (e.g. cardiovascular) over the entire ICU stay. Max SOFA always referred to the highest total SOFA score measured in a prespecified time interval, and Mean SOFA was always calculated by taking the average of

all total SOFA scores in the prespecified time interval. These intervals varied in length, but generally they were equal to the complete ICU stay. Definitions of Delta SOFA were not consistent. Generally, Delta SOFA was defined as Total Max − Admission SOFA [4, 7, 11, 14, 18, 20, 21], but some studies used different definitions [7, 17, 19]. Modified SOFA scores were adapted SOFA scores (e.g. by using a surrogate of the Glasgow Coma Scale).

Best AUCs were found for Max SOFA (range = 0.792 to 0.922) and Total Max SOFA (range = 0.69 to 0.921), while the lowest AUC was found for Delta SOFA (range = 0.51 to 0.828). P-values of HL-statistics ranged from 0.33 to 0.95 for Total Max SOFA and were all beneath 0.05, indicating poor fit, for Delta SOFA and Mean SOFA.

Performance of individual components of SOFA

Four studies (22%) evaluated individual components of SOFA [10, 14, 16, 21] (Table 5). The cardiovascular component performed best in [21] and the neurological component in [10], while the hepatic component did worst in both [10] and [21]. In [16], the Max cardiovascular component had a higher AUC than the other derivatives of the cardiovascular component.

Studies comparing derivatives of SOFA with similar derivatives of the Logistic Organ Dysfunction System (LODS) score and/or the Multiple Organ Dysfunction Score (MODS) found good, comparable discrimination, showing a similar pattern of performance of the different derivatives [10, 14–17]. In one study, however, all derivatives of the cardiovascular component of SOFA did better than that of MODS [16].

Performance of SOFA combined with other models and/or covariates

Six studies (33%) evaluated SOFA combined with other models and covariates [4–7] (Table 6) and [8, 9] (Table 7).

One study compared the APACHE II model alone to APACHE II combined with each one of Total Max SOFA, Delta SOFA and Admission SOFA [4]. Overall performance and discrimination were both improved by the addition of Total Max SOFA and of the Delta SOFA, especially in emergency ICU admissions.

Three studies compared the Simplified Acute Physiology Score (SAPS) II model to the SAPS II model when combined with additional information [5, 8, 9]. One study found that the discriminative ability of SAPS II could be improved by combining it with Mean and Max SOFA scores, event information, and diagnosis information [5]. Two studies built temporal SOFA models and are described in the next section [8, 9].

Two studies combined SOFA with other covariates [6, 7]. The first study evaluated different combinations of SOFA derivatives and age [6]. Highest discriminative ability (AUC = 0.807) was found with the combination of age, Min SOFA, Max SOFA and SOFA trend (using the categories increased, unchanged, and decreased) over 5 days. The second study compared a model based on Max SOFA alone with a model including Max SOFA and infection, and a model including Max SOFA, infection and age [7]. The last model had very good calibration and discrimination, and outperformed the model based on Max SOFA alone.

Performance of temporal SOFA models using pattern discovery

Two studies (11%) of the same research group used pattern discovery to develop temporal models including SAPS II and SOFA data [8,9] (Table 7). The first study used a data-driven algorithm to discover frequent sequences of SOFA scores, categorized as low, medium and high [8]. On all days examined (the first 5 days) the temporal SAPS II model including the frequent SOFA patterns (called episodes) had better accuracy, indicated by lower Brier scores, than the original model. On days 2, 4 and 5 these differences were statistically significant. In the second study the same algorithm was used to discover frequent patterns of individual organ failure (OF) scores (categorized as failure or non-failure) [9] for days 2 to 7. A temporal SAPS II model including the frequent OF patterns was compared to the original (recalibrated) model, the temporal SAPS II model described in [8] and a temporal SAPS II model including a weighted average of the SOFA scores. Except for day 7 the model including frequent OF patterns performed best in terms of both discrimination and accuracy as measured by the Brier score [9].

Discussion

To our knowledge this is the first systematic review on the use of SOFA-based models to predict the risk of mortality in ICU patients. In this review, we show that although the 18 identified studies all focused on evaluating a SOFA-based score or model in predicting mortality they widely differed in the SOFA derivatives used, the time after admission on which the prediction was made, the outcome (hospital or ICU mortality), the prognostic performance measures considered, the way a study was reported, and the way the models were validated. This hampers the quantitative comparability of study results. Despite the fact that most studies scored well on most methodological quality dimensions, model validation still formed a weak spot: in some studies there was no report on how performance measures were obtained and in others there was no independent validation set used. The AUC of SOFA-based models was good to very good and

did not lag much behind APACHE II/III and was competitive with a SAPS II model. When reported, the Hosmer-Lemeshow tests did not indicate poor fit (i.e. there were no significant departures between the predicted probabilities and the respective observed mortality proportions). Model with sequential SOFA seem to have comparable performance to other organ failure scores. Combining SOFA-based derivatives with admission severity of illness scores clearly improved predictions.

Among the used SOFA derivatives are the SOFA score on admission, Maximum SOFA score over the entire ICU stay or the sum of highest SOFA components over ICU stay. Only ten studies reported on the use of SOFA derivatives as covariates in a logistic regression model, the other eight studies did not use models or did not report on such use. The score itself, without using a probabilistic model would allow for obtaining an AUC representing the likelihood that a non-surviving patient would have a higher SOFA score than a patient that would survive. As the SOFA score itself does not give a quantitative estimation of the risk of mortality, calibration and accuracy cannot be assessed for the SOFA score itself. Remarkably, only five of the ten studies fitting a logistic regression model reported on the use of an independent data set to validate the model. Due to these differences in the use of SOFA scores and in the methodological approach and quality, results of individual studies are very hard to compare and meta-analyse.

Most studies evaluated prognosis based on SOFA scores in the first 24 hours after ICU admission. Good to excellent discrimination between survivors and non- survivors were reported, which did not markedly differ from that of traditional models such as APACHE II or SAPS II. This relatively good performance of SOFA is remarkable, given the fact that SOFA is based on fewer physiologic parameters and that it does not include information on reason for admission or co-morbidity. On the other hand, information on instituted treatments, such as vasopressors and mechanical ventilation is included in SOFA but not in APACHE II or SAPS II. We would like to stress that SAPS and APACHE models were developed for predicting hospital mortality, hence when comparing SOFA-based models to this family of admission-based models it is more appropriate to use hospital mortality rather than ICU mortality as the outcome. Table 1 shows that this design principle was not always followed.

It can be expected that adding information on the course of the ICU treatment, as reflected by sequential SOFA scores, will improve the accuracy of predicting the likelihood of survival. Indeed, studies that

evaluated the prognostic value of highest SOFA scores during ICU stay found excellent discrimination as reflected in high AUCs. It should be stressed, however, that most severe organ failure and highest SOFA scores might well be found just prior to death. The clinical relevance of predicting a high likelihood of dying just before actual death is limited. Interestingly, the one study that evaluated Max SOFA over the first 5 days of admission instead of over the entire ICU stay found an AUC of 0.79, almost the same as the AUC for a single SOFA-score at admission [17].

A high Delta SOFA indicates increasing organ dysfunction during ICU stay, it was expected to be highly predictive of mortality. In contrast, discrimination of survivors from non-survivors by Delta SOFA alone appeared to be poor. This may be explained by the fact that Delta SOFA may be relatively low in patients with already a very high SOFA score at admission. Furthermore, Delta SOFA does not take into account whether organ functioning improves after the SOFA score reaches a peak value.

Combining information of severity of illness at admission and information on the course of illness during treatment, in contrast to comparing them, seems promising and two strategies have been adopted. In the first strategy a prognostic model at admission was combined with a pre-specified SOFA derivative such as Delta SOFA or Max SOFA. Indeed, in our review we found that the studies combining Delta SOFA or Max SOFA with APACHE II or SAPS II reported on better discrimination between survivors and non-survivors for the combined models than for either APACHE II or SAPS II alone [4, 5]. A second strategy is to combine severity of admission scores with data-driven patterns of SOFA or individual organ failure scores (e.g. "two days of renal failure accompanied with recovery of the renal system") instead of using pre-specified SOFA derivatives. Two studies adopted this strategy and showed that models based on SAPS-II and temporal patterns outperformed models based on the SAPS II alone [8, 9].

Conclusions

Interest in models based on the score, introduced a decade ago, is increasing in recent years. Although the heterogeneity of published studies hampers drawing precise conclusions about the optimal derivatives of SOFA scores, the following general conclusions may be drawn. Models based on SOFA scores at admission seem to be competitive with severity of illness models limited to the first 24 hours of admission. Performance of models based on sequential SOFA scores is comparable to that of other organ failure scores. Based on current evidence we advocate the combination of a traditional model based on data from the first 24 hours after ICU admission (e.g. APACHE IV) with sequential SOFA scores (e.g. maximum SOFA or a SOFA score pattern over a specified time interval). Such a model should be validated in a large

independent dataset.

Key messages

- SOFA-based models evaluated on their prognostic performance fell under the categories: single SOFA scores at fixed times; sequential SOFA measurements; individual SOFA components; combination of SOFA with other covariates; and SOFA patterns automatically discovered from the data.
- For predicting mortality SOFA-based models at admission seem to be competitive with severity of illness models limited to the first 24 hours of admission and models based on sequential SOFA scores have comparable performance to other organ failure scores.
- The combination of SOFA-based models with admission-based models results in superior prognostic performance than each model alone.
- Studies should use an independent validation set to assess performance and should apply multiple performance measures preferably covering discrimination, calibration, and accuracy.

Abbreviations

APACHE = Acute Physiology And Chronic Health Condition; AUC = Area Under the Receiver Operating; Characteristic Curve; HL statistics = Hosmer-Lemeshow statistics; ICU = Intensive Care Unit; LODS = Logistic Organ Dysfunction System; MODS = Multiple Organ Dysfunction Score; OF scores = Organ Failure scores; SAPS = Simplified Acute Physiology Score; SOFA = Sequential Organ Failure Assessment.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

LM carried out the search queries, reviewed the articles, assessed their quality and drafted the paper. AAH conceived of the study, reviewed the articles and participated in its design and coordination and helped to draft the manuscript. EdJ assessed the quality of the studies and participated in its design and coordination and helped to draft the manuscript All authors read and approved the final manuscript.

References

1. Vincent J, De Mendonça A, Cantraine F, Moreno R, Takala J, Suter P, Sprung C: **Use of the SOFA score to assess the incidence of organ dysfunction/failure in intensive care units: Results of a multicenter, prospective study**. *Critical Care Medicine* 2000, **26**:1793–1800.
2. Vincent J, Ferreira F, Moreno R: **Scoring systems for assessing organ dysfunction and survival**. *Crit Care Clinics* 2000, **16**:353–366.
3. Hayden J, Côté P, Bombardier C: **Evaluation of the Quality of Prognosis Studies in Systematic Reviews**. *Ann of Intern Med* 2006, **144**:427–437.
4. Ho K: **Combining Sequential Organ Failure Assessment (SOFA) score with Acute Physiology and Chronic Health Evaluation (APACHE) II score to predict hospital mortality of critically ill patients**. *Anaesthesia and Intensive Care* 2007, **35**:515–521.
5. Rivera-Fernández R, Nap R, Vázquez-Mata G, Miranda D: **Analysis of physiologic alterations in intensive care unit patients and their relationship with mortality**. *Journal of Critical Care* 2007, **22**:120–128.
6. Cabré L, Mancebo J, Solsona J, Saura P, Gich I, Blanch L: **Multicenter study of the multiple organ dysfunction syndrome in intensive care units: The usefulness of sequential organ failure assessment scores in decision making**. *Intensive Care Medicine* 2005, **31**:927–933.
7. Kajdacsy-BallaAmaral A, Andrade F, Moreno R, Artigas A, Cantraine F, Vincent J: **Use of the sequential organ failure assessment score as a severity score**. *Intensive Care Medicine* 2005, **31**:243–249.
8. Toma T, Abu-Hanna A, Bosman RJ: **Discovery and inclusion of SOFA score episodes in mortality prediction**. *J Biomed Inform* 2007, **40**:649–660.
9. Toma T, Abu-Hanna A, Bosman R: **Discovery and integration of univariate patterns from daily individual organ-failure scores for intensive care mortality prediction**. *Artificial Intelligence in Medicine* 2008, **43**:47–60.
10. Khwannimit B: **A comparison of three organ dysfunction scores: MODS, SOFA and LOD for predicting ICU mortality in critically ill patients**. *Journal of the Medical Association of Thailand* 2007, **90**:1074–1081.
11. Ho K, Lee K, Williams T, Finn J, Knuiman M, Webb S: **Comparison of acute physiology and chronic health evaluation (APACHE) II score with organ failure scores to predict hospital mortality**. *Anaesthesia* 2007, **62**:466–473.
12. Holtfreter B, Bandt C, Kuhn S, Grunwald U, Lehman C, Schütt C: **Serum osmolality and outcome in intensive care unit patients**. *Acta Anaesthesiologica Scandinavica* 2006, **50**:970–977.
13. Gosling P, Czyz J, Nightingale P, Manji M: **Microalbuminuria in the intensive care unit: Clinical correlates and association with outcomes in 431 patients**. *Critical Care Medicine* 2006, **34**:2158–2166.
14. Zygun D, Laupland K, Fick G, Sandham J, Doig C, Chu Y: **Limited ability of SOFA and MOD scores to discriminate outcome: A prospective evaluation in 1,436 patients**. *Canadian Journal of Anesthesia* 2005, **52**:302–308.
15. Timsit J, Fosse J, Troché G, DeLassence A, Alberti C, Garrouste-Orgeas M: **Calibration and discrimination by daily logistic organ dysfunction scoring comparatively with daily sequential organ failure assessment scoring for predicting hospital mortality in critically ill patients**. *Critical Care Medicine* 2002, **30**:2003–2013.
16. Peres Bota D, Melot C, Lopes Ferreira F, Ba V, Vincent J: **The multiple organ dysfunction score (MODS) versus the sequential organ failure assessment (SOFA) score in outcome prediction**. *Intensive Care Medicine* 2002, **28**:1619–1624.
17. Pettilä V, Pettilä M, Sarna S, Voutilainen P, Takkunen O: **Comparison of multiple organ dysfunction scores in the prediction of hospital mortality in the critically ill**. *Critical Care Medicine* 2002, **30**:1705–1711.
18. Junger A, Engel J, Benson M, Böttger S, Grabow C, Hartmann B: **Discriminative power on mortality of a modified sequential organ failure assessment score for complete automatic computation in an operative intensive care unit**. *Critical Care Medicine* 2002, **30**:338–342.

19. Lopes Ferreira F, Peres Bota D, Bross A, Mélot C, Vincent J: **Serial evaluation of the SOFA score to predict outcome in critically ill patients**. *Journal of the American Medical Association* 2001, **286**:1754–1758.

20. Janssens U, Graf J, Radke P, Königs B, Koch K: **Evaluation of the sofa score: A single-center experience of a medical intensive care unit 303 consecutive patients with predominantly cardiovascular disorders**. *Intensive Care Medicine* 2001, **26**:1037–1045.

21. Moreno R, Vincent J, Matos R, Mendonça A, Cantraine F, Thijs L, Takala J, Sprung C, Antonelli M, Bruining H, Willats S: **The use of maximum SOFA score to quantify organ dysfunction/failure in intensive care. Results of a prospective, multicentre study**. *Intensive Care Medicine* 1999, **25**:686–696.

Table 1. Study Characteristics.

	Study design		Population	Models	Variables		Comparison	
	Setting (Location)[a]	Study period[b]	N[c]/ICU Type[d]/ Mortality%[e]	Model/ Valid.[f]	SOFA Abstractions[g]	Others[h]	Standard Model[i]	Mort.[j]
Toma (2008) [9]	1 ICU (NL)	Jul 98-Aug 05	2928/Mix/ H=24	LR/Ind.	Seq of OF[1]	SAPS II	SAPS II	H
Toma (2007) [8]	1 ICU (NL)	Jul 98-Aug 05	6276/Mix/ H=11	LR/Ind.	Seq of SOFA[2]	SAPS II	SAPS II	H
Ho (2007) [4]	1 multidisc ICU (AU)	Jan 05-Dec 05	1311/Mix/ H=14.5	LR/Ind.	TMS Adm Delta (TMS−Adm)	APACHE II	APACHE II	H
Ho (2007) [11]	1 multidisc ICU (AU)	Jan 05-Dec 05	1311/Mix/ H=14.5	No	TMS Adm Delta (TMS−Adm)	No	APACHE II APS, RPH	H
Holtfreter (2006) [12]	1 ICU (DE)	42 months	933/Mix/ H=25/I=23.9	No	Adm	No	16 variables APACHE II	H
Zygun (2005) [14]	3 ICUs (CA)	May 00-Apr 01	1436/Mix/ H=35.1/I=27	LR/NR	Adm TMS, Mean (ICU stay), Delta (TMS−Adm), Adm (i)	No	MODS	H/ I
Cabré (2005) [6]	79 ICUs (75 ES, 4 L-Am)	Feb 01-Mar 01	1324/Mix/ H=44.6/I=37.3	LR/NR	Min (MODS period), Max (MODS period), 5-day trend[3]	Age	No	H
Timsit (2002) [15]	6 ICUs (FR)	24-months	1685/Mix/ H=30.3/I=22.5	LR/Ind.⋆	D1-7, D1-7 (mod)	No	LODS	H
Pettilä (2002) [17]	1 med-surg ICU (FI)	NR	520/Mix/ H=30/I=16.5	No	Adm, D5, Max (5d), Delta (d5−d1), TMS	No	APACHE III MODS LODS	H
Janssens (2000) [20]	1 med ICU (DE)	Nov 97-Feb 98	303/Med/ H=14.5/I=6.3	LR/NR	Adm, TMS, Delta (TMS−Adm)	No	No	H
Khwannimit (2007) [10]	1 ICU (TH)	Jul 04-Mar 06	1782/Mix/ H=22/I=16.4	No	Adm	No	MODS, SOFA LODS	I
Rivera-Fernández (2007) [5]	55 ICUs (EU)	2 months in 97/98	6409/Mix/ H=20.6/I=13.9	LR/Ind.	Mean (ICU stay), Max (ICU stay)	SAPS II diagnosis events	SAPS II	I
Gosling (2006) [13]	1 general ICU (UK)	Nov 02-Oct 03	431/Mix/ I=20.9	No	Adm SOFA	No	APACHE II urine albumin and 5 other factors	I
Kajdacsy-Balla Amaral (2005) [7]	40 ICUs (1 AU, 35 EU, 1 N-Am, 3 S-Am)	May 1 95-May 31 95	748 (6 countries)/ Mix/I=21.5	LR/NR	Adm, TMS, Delta (48h−Adm), Delta (TMS−Adm)	Different parameters	No	I
Junger (2002) [18]	1 operative ICU (DE)	Apr 99-Mar 00	524/Surg/ I=12.4	No	Max (ICU stay), TMS, Delta (TMS−Adm), Adm (mod)	No	No	I
Ferreira (2001) [19]	1 med-surg ICU (BE)	Apr 99-Jul 99	352/Mix/ I=23	No	Adm, 48h, 96h, Delta (48h−Adm), Delta (96h−Adm), Max (ICU stay), Mean (ICU stay), Total	No	No	I
Moreno (1999) [21]	40 ICUs (1 AU, 35 EU, 1 N-Am, 3 S-Am)	May 95	1449/Mix/ H=26/I=22	LR/NR	Adm, TMS, Delta (TMS−Adm), Adm (i)	No	No	I
Bota (2002) [16]	1 ICU (BE)	Apr-Jul99 Oct-Nov99 Jul-Sep00	949/Mix/ 29.1	No	Adm, 48h, 96h, Dis, Max (24h), Adm (c), 48h (c), 96h (c), Dis (c), Max (c, 24h)	No	APACHE II MODS	U

[a]: ICU=Intensive Care Unit, multidisc=multidisciplinary, med=medical, surg=surgical, NL=The Netherlands, TH=Thailand, AU=Australia, EU=European Union, NO=Norway, DE=Germany, UK=United Kingdom, CA=Canada, ES=Spain,

L-Am=Latin-America, N-Am=North-America, S-Am=South-America, FR=France, BE=Belgium, FI=Finland.
[b]: NR=Not Reported.
[c]: N=Number of patients.
[d]: Mix=Mixed, Med=medical, Surg=surgical.
[e]: H=Hospital mortality, I=ICU mortality, M=Undefined mortality.
[f]: Model=Model type reported, Valid.=Validation method, LR=Logistic Regression, Ind.=Independent validation set used (* indicates the use of bootstrapping), No=No model was used, NR=Not Reported.
[g]: seq=sequences, OF=individual Organ Failure scores, SOFA=Sequential Organ Failure Assessment, Adm=Admission, Dis=Discharge, Max=Maximum, TMS=Total Maximum SOFA, cust=customized, mod=modified, i=individual components of SOFA, c=cardiovascular component of SOFA, Dx=Day x (x=day number), xd=x days (x=number of days), xh=x hours (x=number of hours), 1=Sequences of categorized individual components of SOFA (Failure-Non failure), 2=Sequences of categorized SOFA scores (High-Medium-Low), 3=SOFA trend over 5 days (-1 if SOFA is decreased, 0 if SOFA is unchanged, 1 if SOFA is increased).
[h]: SAPS=Simplified Acute Physiology Score, APACHE=Acute Physiology And Chronic Health Evaluation.
[i]: SAPS=Simplified Acute Physiology Score, APACHE=Acute Physiology And Chronic Health Evaluation, APS=Acute Physiology Score, SOFA=Sequential Organ Failure Assessment, LODS=Logistic Organ Dysfunction System, MODS=Multiple Organ Dysfunction Score, RPH= Royal Perth Hospital Intensive Care Unit.
[j]: Mort.=Mortality, H=Hospital mortality, I=ICU mortality, U=Undefined mortality.

Table 2: Quality score of included studies.

	Study participation max 8 pts	Prognostic factor max 3 pts	Outcome measurement max 1 pt	Analysis max 8 pts	Total score max 20 pts
Toma (2008) [9]	8	3	1	7.5	19
Toma (2007) [8]	8	2.5	1	8	19.5
Khwannimit (2007) [10]	8	1	1	3.5	13.5
Ho (2007) [4]	8	3	1	7	19
Ho (2007) [11]	8	2	1	5	16
Rivera-Fernandez (2007) [5]	7	1	1	7.5	16.5
Holtfreter (2006) [12]	8	1.5	1	5	15.5
Gosling (2006) [13]	8	1.5	1	4	14.5
Zygun (2005) [14]	8	2	1	5.5	16.5
Cabre (2005) [6]	8	2	1	4	15
Kajdacsy-Balla Amaral (2005) [7]	8	3	1	5	17
Timsit (2002) [15]	8	2.5	1	7.5	19
Bota (2002) [16]	7.5	1	0	3	11.5
Pettila (2002) [17]	8	1	1	7.5	17.5
Junger (2002) [18]	7	2	1	3	13
Ferreira (2001) [19]	8	2.5	1	3	14.5
Janssens (2000) [20]	8	2	1	3.5	14.5
Moreno (1999) [21]	8	2.5	1	3.5	15

max=maximum, criteria for quality assessment are based on a 20 item list [see Additional data file 1].

Table 3. Performance at admission or a fixed time thereafter.

Admission SOFA	AUC	Brier	H/C-statistics	Compared to	AUC	Brier	H/C-statistics	Mort.
Ho (2007) [11]	0.791	0.1	C=7.97 p=0.437	APACHE II	0.858	0.09		H
				APS	0.829	0.09	C=2.9 p=0.890	H
				RPHICU	0.822	0.09	C=4.7 p=0.198	H
Holtfreter (2006) [12]	0.72			APACHE II	0.785			H
Zygun (2005) [14]	0.67		H/C=8.8 p=0.38	MODS	0.62		H/C=10.28 p=0.17	H
Timsit (2002) [15]	0.72		H/C=4.55 p=0.8	LODS	0.726		H/C=10.4 p=0.16	H
Pettilä (2002) [17]	0.776			APACHE III	0.825			H
				LODS	0.805			H
				MODS	0.695			H
Khwannimit (2007) [10]	0.8786			LODS	0.8802			H
				MODS	0.8606			I
Gosling (2006) [13]	0.61			APACHE II	0,62			I
Zygun (2005) [14]	0.67		H/C=11.66 p=0.17	MODS	0.63		H/C=14.29 p=0.05	I
Moreno (1999) [21]	0.772							I
Bota (2002) [16]	0.872			APACHE II	0.88			U
				MODS	0.856			U
Ferreira (2001) [19]	0.79							I
Janssens (2000) [20]	0.82			SAPS II	0.77			H
Other scoring moments	AUC	Brier	H/C-statistics	Compared to	AUC	Brier	H/C-statistics	Mort.
Bota (2002) [16] 48 hours	0.844			MODS	0.834			U
Ferreira (2001) [19] 48 hours	0.78							I
Bota (2002) [16] 96 hours	0.847			MODS	0.861			U
Ferreira (2001) [19] 96 hours	0.82							I
Timsit (2002) [15], day 2	0.742		H/C=11.1 p=0,2	LODS	0.742			H
Timsit (2002) [15], day 3	0.762		H/C=9.94 p=0.27	LODS	0.762			H
Timsit (2002) [15], day 4	0.766		H/C=10.5 p=0.23	LODS	0.766			H
Timsit (2002) [15], day 5	0.746		H/C=13.6 p=0.09	LODS	0.746			H
Pettilä (2002) [17], day 5	0.727			LODS	0.76			H
				MODS	0.744			H
Timsit (2002) [15], day 6	0.763		H/C=12.2 p=0.14	LODS	0.763			H
Timsit (2002) [15], day 7	0.746			LODS	0.764			H
Bota (2002) [16], final	0.897			MODS	0.869			H

Mort.=Mortality, H=Hospital, U=Undefined, I=Intensive Care Unit, APACHE=Acute Physiology and Chronic Health Evaluation, APS=Acute Physiology Score (APACHE without chronic health and age condition), SAPS=Simplified Acute Physiology Score, SOFA=Sequential Organ Failure Assessment, LODS=Logistic Organ Dysfunction System, MODS=Multiple Organ Dysfunction Score, RPHICU= Royal Perth Hosp. Intensive Care Unit, AUC=Area Under the Receiver Operating Curve, H/C=H- or C- statistic (undefined).

Table 4. Performance for sequential SOFA.

Max SOFA	AUC	Brier	H/C-statistics	Comp.	AUC	H/C-statistics	Mort.
Pettilä (2002) [17], 5 days	0.792			LODS	0.827		H
				MODS	0.795		H
Junger (2002) [18], ICU stay	0.922						I
Bota (2002) [16], 24hrs period	0.898			MODS	0.9		U
Ferreira (2001) [19], ICU stay	0.9						I
Total Max SOFA	AUC	Brier	H/C-statistics	Comp	AUC	H/C-statistics	Mort.
Ho (2007) [11], ICU stay	0.829	0.1	C=7.4 p=0.496				H
Zygun (2005) [14], ICU stay	0.7		9.2 p=0.33	MODS	0.65	8.07 p=0.43	H
Pettilä (2002) [17], ICU stay	0.816			LODS	0.839		H
				MODS	0.817		H
Zygun (2005) [14], ICU stay	0.69		7.30 p=0.50	MODS	0.64	9.09 p=0.33	I
Kajdacsy-Balla	0.84		H: p=0.95 C: p=0.54				I
Amaral (2005) [7], ICU stay							
Junger (2002) [18], ICU stay	0.921						I
Moreno (1999) [21], ICU stay	0.847						I
Janssens (2000) [20], ICU stay	0.86						H
Delta SOFA	AUC	Brier	H/C-statistics	Comp	AUC	H/C-statistics	Mort.
Ho (2007) [11], TMS − Adm	0.635	0.12	C=20.2 p=0.001				H
Zygun (2005) [14], TMS − Adm	0.54		53.48 p<0.01	MODS	0.55	31.2 p<0.01	H
Pettilä (2002) [17], day 5 − Adm	0.6			LODS	0.633		H
				MODS	0.653		H
Zygun (2005) [14], TMS − Adm	0.51		98.01 p<0.01	MODS	0.52	70.52 p<0.01	I
Junger (2002) [18], TMS − Adm	0.828						I
Moreno (1999) [21], TMS − Adm	0.742						I
Ferreira (2001) [19], 48hrs − Adm	0.69						I
Ferreira (2001) [19], 96hrs − Adm	0.62						I
Janssens (2000) [20], TMS − Adm	0.62						H
Mean SOFA	AUC	Brier	H/C-statistics	Comp	AUC	H/C-statistics	Mort.
Zygun (2005) [14], ICU stay	0.77		22.66 p<0.01	MODS	0.74	46.13 p<0.01	H
Zygun (2005) [14], ICU stay	0.79		28.92 p<0.01	MODS	0.75	42.72 p<0.01	I
Ferreira (2001) [19], ICU stay	0.88						I
Total SOFA	AUC	Brier	H/C-statistics	Comp	AUC	H/C-statistics	Mort.
Ferreira (2001) [19], ICU stay	0.85						I
Modified SOFA	AUC	Brier	H/C-statistics	Comp	AUC	H/C-statistics	Mort.
Timsit (2002) [15], Adm	0.729		11 p=0.2	LODS	0.733	11.3 p=0.19	H
Timsit (2002) [15], day 2	0.752		8.3 p=0.4	LODS	0.748		H
Timsit (2002) [15], day 3	0.773		11.3 p=0.19	LODS	0.761		H
Timsit (2002) [15], day 4	0.779		7.3 p=0.5	LODS	0.76		H
Timsit (2002) [15], day 5	0.763		14.4 p=0.07	LODS	0.749		H
Timsit (2002) [15], day 6	0.784		11 p=0.17	LODS	0.79		H
Timsit (2002) [15], day 7	0.768		6.3 p=0.62	LODS	0.746		H
Junger (2002) [18], Adm	0.799						I

Comp.=Compared to, Mort.=Mortality, H=Hospital, I=ICU=Intensive Care Unit, U=Undefined, SOFA=Sequential Organ Failure Assessment, LODS=Logistic Organ Dysfunction System, MODS=Multiple Organ Dysfunction Score, AUC=Area Under the Receiver Operating Curve, H/C=H- or C- statistic (undefined), max=maximum, Adm=admission, TMS =total max SOFA (always measured over entire ICU stay), hrs=hours.

Table 5. Performance for individual components of SOFA.

Cardiovascular SOFA	AUC	Compared to	AUC	Mortality
Zygun (2005) [14], Adm	0.68	MODS	0.63	Hospital
Khwannimit (2007) [10], Adm	0.725	LODS	0.772	ICU
		MODS	0.726	ICU
Zygun (2005) [14], Adm	0.74	MODS	0.64	ICU
Moreno (1999) [21], Adm	0.802			ICU
Bota (2002) [16], Adm	0.75	MODS	0.694	Undefined
Bota (2002) [16], 48 hours	0.732	MODS	0.675	Undefined
Bota (2002) [16], 96 hours	0.739	MODS	0.674	Undefined
Bota (2002) [16], discharge	0.781	MODS	0.75	Undefined
Bota (2002) [16], max	0.821	MODS	0.75	Undefined
Respiratory SOFA	AUC	Compared to	AUC	Mortality
Khwannimit (2007) [10], Adm	0.725	LODS	0.704	ICU
		MODS	0.71	ICU
Moreno (1999) [21], Adm	0.736			ICU
Hepatic SOFA	AUC	Compared to	AUC	Mortality
Khwannimit (2007) [10], Adm	0.539	LODS	0.563	ICU
		MODS	0.539	ICU
Moreno (1999) [21], Adm	0.655			ICU
Renal SOFA	AUC	Compared to	AUC	Mortality
Khwannimit (2007) [10], Adm	0.678	LODS	0.727	ICU
		MODS	0.659	ICU
Moreno (1999) [21], Adm	0.739			ICU
Neurological SOFA	AUC	Compared to	AUC	Mortality
Khwannimit (2007) [10], Adm	0.84	LODS	0.822	ICU
		MODS	0.839	ICU
Moreno (1999) [21], Adm	0.727			ICU
Coagulation SOFA	AUC	Compared to	AUC	Mortality
Khwannimit (2007) [10], Adm	0.623	LODS	0.59	ICU
		MODS	0.632	ICU
Moreno (1999) [21], Adm	0.684			ICU

ICU=Intensive Care Unit, SOFA=Sequential Organ Failure Assessment, LODS=Logistic Organ Dysfunction System, MODS=Multiple Organ Dysfunction Score, AUC=Area Under the Receiver Operating Curve, max=maximum, Adm=admission

Table 6. Performance for combined models.

APACHE II	Given by	AUC	Brier	H/C statistics	Mortality
APACHE II	Ho (2007) [4]	0.859	0.09	C=10 p=0.189	Hospital
APACHE II + Total Max SOFA	Ho (2007) [4]	0.875	0.086	C=10.1 p=0.261	Hospital
APACHE II + Delta SOFA	Ho (2007) [4]	0.874	0.086	C=7.5 p=0.485	Hospital
APACHE II + Admission SOFA	Ho (2007) [4]	0.861	0.09	C=9.3 p=0.318	Hospital
SAPS II	Given by	AUC	Brier	H/C statistics	Mortality
SAPS II	R-F (2007) [5]	0.8		H: 12.02 p > 0.05	ICU
SAPS II + Diagnosis	R-F (2007) [5]	0.84			ICU
SAPS II + Diagnosis + Events	R-F (2007) [5]	0.91			ICU
SAPS II + Mean SOFA + Max SOFA + Events	R-F (2007) [5]	0.93			ICU
SAPS II + Mean SOFA + Max SOFA + Events + Diagnosis	R-F (2007) [5]	0.95		H: 12.02 p > 0.05	ICU
Other covariates	Given by	AUC	Brier	H/C statistics	Mortality
Min SOFA + Max SOFA + SOFA trend over 5 days + Age	Cabré (2005) [6]	0.807			Hospital
Max SOFA > 13 + Min SOFA > 10 + Positive SOFA trend + Age > 60	Cabré (2005) [6]	0.750			Hospital
Max SOFA > 10 + Min SOFA > 10 + Positive SOFA trend + Age > 60	Cabré (2005) [6]	0.758			Hospital
Total Max SOFA	K-BA (2005) [7]	0.841			ICU
Total Max SOFA + Infection	K-BA (2005) [7]	0.845			ICU
Total Max SOFA + Infection + Age	K-BA (2005) [7]	0.853		C: p=0.37 H: p=0.73	ICU

ICU=Intensive Care Unit, APACHE=Acute Physiology and Chronic Health Evaluation, SAPS=Simplified Acute Physiology Score, SOFA=Sequential Organ Failure Assessment, AUC=Area Under the Receiver Operating Curve, min=minimum, max=maximum, RF=Rivera-Fernández, K-BA=Kajdacsy-Balla Amaral.

Table 7. Performance for temporal models using pattern discovery.

		Brier score						
SAPS II + SOFA	Given by	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7
Recalibrated SAPS II	Toma (2007) [8]	0.059	0.132	0.17	0.18	0.182		
Recalibrated SAPS II	Toma (2008) [9]		0.175	0.168	0.198	0.199	0.215	0.23
Temporal SOFA model	Toma (2007) [8]	0.058	0.128	0.161	0.171	0.166		
Temporal SOFA model	Toma (2008) [9]		0.168	0.17	0.195	0.183	0.206	0.211
Temporal wSOFA model	Toma (2008) [9]		0.166	0.175	0.199	0.19	0.21	0.224
Temporal OF model	Toma (2008) [9]		0.161	0.166	0.187	0.175	0.195	0.216
					AUC			
SAPS II + SOFA	Given by	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7
Recalibrated SAPS II	Toma (2008) [9]		0.761	0.746	0.692	0.66	0.643	0.645
Temporal SOFA model	Toma (2008) [9]		0.786	0.780	0.713	0.737	0.690	0.722
Temporal wSOFA model	Toma (2008) [9]		0.794	0.771	0.699	0.709	0.672	0.664
Temporal OF model	Toma (2008) [9]		0.794	0.785	0.727	0.740	0.738	0.715

SAPS=Simplified Acute Physiology Score, SOFA=Sequential Organ Failure Assessment, wSOFA=weighted SOFA, OF=Organ Failure, AUC=Area Under the Receiver Operating Curve

Figure Legends

Figure 1. Search flow chart. N = Number of studies.

Additional Files

The following additional data are available with the online version of this paper. Additional data file 1 describes the 20 items of the quality assessment framework.

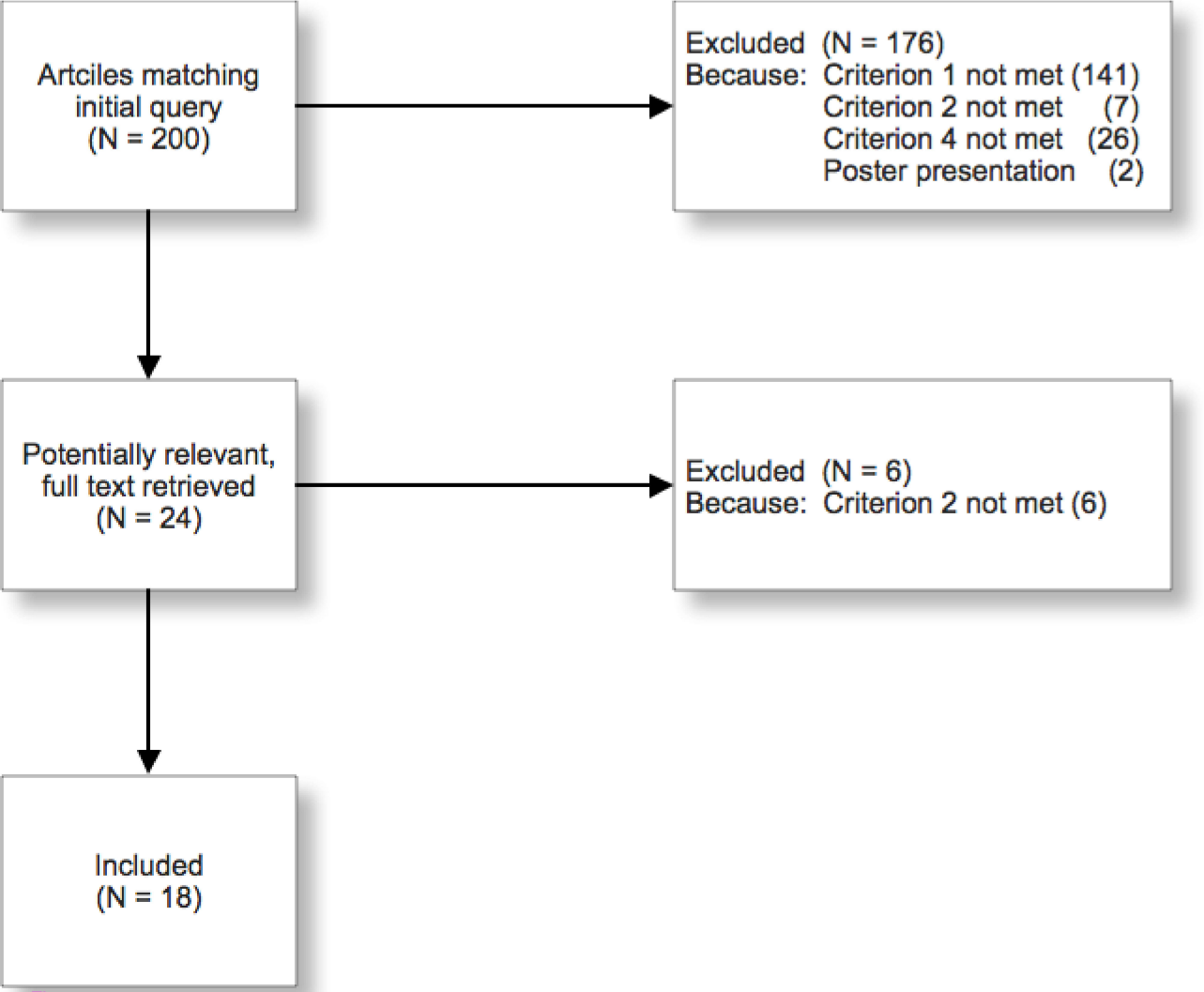
Artciles matching initial query (N = 200)
Excluded (N = 176)
Because: Criterion 1 not met (141)
Criterion 2 not met (7)
Criterion 4 not met (26)
Poster presentation (2)
Potentially relevant, full text retrieved (N = 24)
Excluded (N = 6)
Because: Criterion 2 not met (6)
Included (N = 18)

Figure 1

Additional files provided with this submission:

Additional file 1: ccappendix.pdf, 23K
http://ccforum.com/imedia/6805080322427807/supp1.pdf