Study protocol v. 1.0

Systematic review of the Sequential Organ Failure Assessment score as a surrogate endpoint in randomized controlled trials

Harm-Jan de Grooth, Jean-Jacques Parienti, [to be determined], Heleen Oudemans-van Straaten

Introduction

The Sequential Organ Failure Assessment (SOFA) score was developed to describe multiple organ dysfunction in the Intensive Care Unit (ICU) on a scale that is easily parameterized [1]. The SOFA score was quickly recognized as a potential surrogate endpoint for randomized controlled trials (RCT's) because serially measured SOFA scores were associated with mortality independent of admission score [2,3]. Because of its scalar nature, demonstrating a treatment effect on SOFA score often requires a smaller sample size than demonstrating a treatment effect on mortality.

Multiple large observational studies have confirmed that serial SOFA derivatives such as delta SOFA, total maximum SOFA and mean SOFA are reliable predictors of mortality, sometimes performing only slightly worse than more sophisticated scoring systems such as APACHE II/III (although in some studies the Area Under the receiver operating characteristics Curve (AUC) is as low as 0.5 or 0.6) [4]. This has led to an increasing popularity of the use of SOFA derivatives as primary or secondary endpoints in RCT's. But the association between serial SOFA scores and mortality cannot be directly carried over from observational studies to RCT's.

Firstly, the *responsiveness* of the SOFA score to intervention-induced changing underlying mortality risk has not been quantified. It is unclear how the SOFA score changes in response to a treatment that changes the underlying risk of mortality. Secondly, the *consistency* of the SOFA score to reflect changes in underlying mortality risk has not been quantified. Even if true mortality-modifying treatments effects are reflected in the SOFA score on average, the validity of the SOFA score as an endpoint could still be doubtful if this relation is inconsistent. Thirdly, it is unclear which derivative of SOFA score (absolute score day X, delta day-X, maximum delta, total maximum, sum over days) is the most responsive and consistent surrogate endpoint. Different derivative scores are currently used in many different RCT's, often without explicit justification.

The aim of this study is to quantify the responsiveness and the consistency of different SOFA derivatives to reflect intervention-related changes in mortality risk. We will use data from a large number of published RCT's that report both SOFA and mortality endpoints. Furthermore, we will try to identify which derivative of the SOFA score is most responsive and consistent in detecting mortality-modifying treatment effects. We will also discuss pitfalls associated with the use of surrogate endpoints in general and we will highlight the need for thorough validation [5–7].

Potential impact of this study

The SOFA score is gaining popularity as an endpoint in RCT's (figure 2). The results from this study will aid decision makers in the interpretation of the trials that use a SOFA endpoint and will likely influence the design

of future ICU trials. If the SOFA score is shown to be an inconsistent surrogate for mortality, then a treatment-induced decrease in SOFA score cannot be extrapolated to a lower mortality risk. However, if the responsiveness of SOFA score to mortality changes is good and if consistency is acceptable, then the SOFA score may be a reasonable and practical endpoint for many future trials that are not able to recruit the patient numbers required to show a mortality effect.

Methods

Inclusion and search strategy for studies and variables of interest

Eligible for inclusion are RCT's in adult ICU patients reporting both a derivative of SOFA score and a measure of mortality as primary or secondary endpoints. PubMed, MEDLINE and Embase databases are queried using the term *[(sofa OR "sepsis-related organ failure" OR "sepsis related organ failure" OR "sequential organ failure") AND (random* OR RCT)]*. The query will be repeated 1 month before submission. Reports in languages other than English will be excluded from the analysis.

For each RCT, we will register and categorize the trial population category, the intervention being tested, single- or multicenter design, the primary endpoint and the analysis type (intention-to-treat or per-protocol). Trials will be graded according to the Jadad scale [8]. For each treatment arm we will register the sample size, baseline SOFA score, all reported serial SOFA scores (including standard deviation or interquartile ranges and differentiating absolute scores from delta scores) and the reported mortality rates.

Background of surrogate endpoint validation

The statistical validation of surrogate endpoints for clinical trials has been operationalized in various seminal papers and guidelines [5–7,9–11]. For the purpose of our research question, the validation method is limited by several constraints. Firstly, the SOFA score is non-dichotomous and the analysis will be performed at the study level and not at the individual patient level. This precludes a sensitivity-specificity approach to and use of 'proportion explained' types of regression methods[9,11]. Secondly, the included trials do not make up a group of homogeneous interventions but rather represent a common biological pathway of multiple organ dysfunction as a primary determinant of ICU-related mortality. Statistical heterogeneity in the relation between SOFA score and mortality therefore seems inevitable and needs to be modelled explicitly. A final constraint on the statistical method is imposed by the intended readership and the intended applicability of this paper. The results should be usable by nonmathematical *clinicians* who need to make informed decisions based on trials that use SOFA score as an endpoint, and also by *researchers* and methodologists for the design and evaluation of future trials.

Given the above-mentioned goals and constraints, we will use meta-regression as the core analytical method. This allows us to model the odds ratio of mortality as a function of between-group SOFA differences, to weigh each study according to sampling variance and to analyze the residual heterogeneity.

Quantifying responsiveness and consistency

For each trial, mortality will be expressed as the odds ratio (OR) of treatment vs. control group mortality. For studies that report multiple measures of mortality, one measure is chosen in the following order: Mortality measure reported as primary endpoint; 28-day mortality; hospital mortality; 90-day mortality; ICU mortality. For the SOFA score, the unit of analysis for the meta-regression is the *standardized difference* between the control and intervention groups, defined as the between-group SOFA score difference divided by the standard deviation (SD) of the SOFA score (square root of the mean of variances of both groups). The standardized difference is used instead of the absolute difference to normalize the SOFA effect size across trials with different SOFA score distributions. When SOFA score is reported as median and IQR, the median will be used as the best unbiased estimator of the mean and the SD will be approximated as IQR/1.35.

A mixed-effects meta-regression model is used with log(OR) as dependent variable, SOFA score (standardized difference) as fixed effect independent variable and a random intercept for each study. The random intercept per study is applied to model heterogeneity explicitly. Fixed- and mixed-effects models produce identical results in the absence of significant between-study heterogeneity, but the mixed-effects leads to appropriately increased standard errors when significant heterogeneity occurs. Each study is weighed by the inverse of the sampling variance of the mortality OR (a function of mortality rate and sample size). A restricted maximum likelihood (REML) estimator will be used to estimate heterogeneity. Residuals will be checked for normality and the goodness of fit of the log-linear model will be compared to power quadratic and power models.

The *responsiveness* of the surrogate endpoint is measured by the coefficient that determines the slope between the standardized between-group SOFA difference and the between-group mortality OR.

The *consistency* of SOFA score as a surrogate endpoint is measured by *tau* and I^2 [12]. *Tau* measures the standardized residual heterogeneity and I^2 describes the percentage total variability that is unexplained by sampling error (chance). Consistency will be defined as good, moderate or poor for I^2 values of <25%, 25-50% and >50%, respectively [13]. The cause of moderate or poor consistency will be explored by adding study-level explanatory variables (e.g. baseline SOFA and trial characteristics) as regressors in the model.

The regression analyses will be performed in R using the *metafor* package [14]. The code for the entire analysis will be published along with the data as supplementary material.

Comparing responsiveness and consistency

The meta-regression will be performed for each derivative of SOFA score: Early absolute score (day 2, 3, 4), late absolute score (day 5-14), total maximum score, delta day-X minus admission and delta maximum minus admission. A study can recur in multiple analyses if more than one SOFA derivative is reported. For this set of regression analyses, the p-values will be corrected for multiple comparisons using the method described by Hommel [15]. The calculation of the different SOFA derivatives will be tabulated for clarity. The responsiveness (regression coefficient) and the consistency (*tau* and I^2) will be compared between the different SOFA derivatives to evaluate whether any derivative is especially superior or inferior for use as a surrogate endpoint.

Regression coefficients will be compared using t-tests on model coefficients and (pooled) error variances [16]. Nonzero *tau* values will be comparing using F-tests on tau^2.

Because the SOFA score was originally designed to quantify sepsis-related organ failure, a subgroup analysis will be performed in the trials with sepsis patient populations. Responsiveness and consistency parameters will be compared for significant differences (correcting for multiple comparisons).

Figures

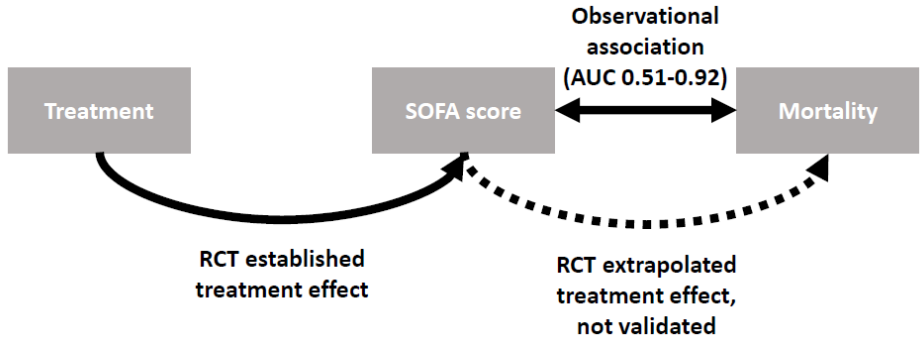

Figure 1. The association between SOFA score and mortality has been demonstrated in observational studies with varying degrees of discrimination (AUC: Area Under the receiver operating characteristic Curve) [4]. The aim of this study is to validate the SOFA score as a surrogate endpoint for mortality in Randomized Controlled Trials (RCT's).

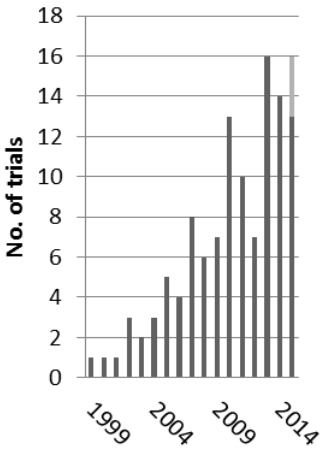

Figure 2. Included studies by year of publication (2015 extrapolated).

References

1. Vincent, J. L. *et al.* The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. *Intensive Care Med.* **22,** 707–10 (1996).

2. Moreno, R. *et al.* The use of maximum SOFA score to quantify organ dysfunction/failure in intensive care. Results of a prospective, multicentre study. Working Group on Sepsis related Problems of the ESICM. *Intensive Care Med.* **25,** 686–96 (1999).

3. Ferreira, F. L., Bota, D. P., Bross, A., Mélot, C. & Vincent, J. L. Serial evaluation of the SOFA score to predict outcome in critically ill patients. *JAMA* **286,** 1754–1758 (2001).

4. Minne, L., Abu-Hanna, A. & de Jonge, E. Evaluation of SOFA-based models for predicting mortality in the ICU: A systematic review. *Crit. Care* **12,** R161 (2008).

5. Prentice, R. L. Surrogate endpoints in clinical trials: Definition and operational criteria. *Stat. Med.* **8,** 431–440 (1989).

6. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. ICH Harmonized Tripartite Guideline: Statistical Principles for Clinical Trials E9. (1998).

7. De Gruttola, V. G. *et al.* Considerations in the evaluation of surrogate endpoints in clinical trials. summary of a National Institutes of Health workshop. *Control. Clin. Trials* **22,** 485–502 (2001).

8. Jadad, A. R. *et al.* Assessing the quality of reports of randomized clinical trials: Is blinding necessary? *Control. Clin. Trials* **17,** 1–12 (1996).

9. Freedman, L. S., Graubard, B. I. & Schatzkin, A. Statistical validation of intermediate endpoints for chronic diseases. *Stat. Med.* **11,** 167–78 (1992).

10. Fleming, T. R. Current issues in non-inferiority trials. *Stat. Med.* **27,** 317–332 (2008).

11. Cleophas, T. J., Zwinderman, A. H. & Chaib, A. H. Novel procedures for validating surrogate endpoints in clinical trials. *Curr. Clin. Pharmacol.* **2,** 123–8 (2007).

12. Higgins, J. P. T., Thompson, S. G., Deeks, J. J. & Altman, D. G. Measuring inconsistency in meta-analyses. *BMJ* **327,** 557–60 (2003).

13. Deeks, J. J., Higgins, J. P. T. & Altman, D. G. in *Cochrane Handbook for Systematic Reviews of Interventions. Version 5.1.0 (updated March 2011).* (eds. Higgins, J. P. T. & Green, S.) (The Cochrane Collaboration, 2011).

14. Viechtbauer, W. Conducting Meta-Analyses in R with the metafor Package. *J. Stat. Softw.* **36,** (2010).

15. Hommel, G. A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* **75,** 383–386 (1988).

16. Andrade, J. M. & Estévez-Pérez, M. G. Statistical comparison of the slopes of two regression lines: A tutorial. *Anal. Chim. Acta* **838,** 1–12 (2014).