Reviewer’s report

**Title:** A comparison of Child-Pugh, APACHE II and APACHE III scoring systems in predicting hospital mortality of patients with liver cirrhosis

**Authors:**

Constantinos Chatzicostas (matrix01gr@yahoo.gr)
Maria Roussomoustakaki (mrousso1@hotmail.com)
Georgios Notas (chief@med.uoc.gr)
Ioannis G Vlachonikolis (i.vlachonikolis@surrey.ac.uk)
Demetrios Samonakis (photein@in.gr)
John Romanos (romanos@med.uoc.gr)
Emmanouel Vardas (evardas@hotmail.com)
Elias A Kouroumalis (kouroum@med.uoc.gr)

**Version:** 1  **Date:** 9 Oct 2002

**Reviewer:** Dr Bekele Afessa

**Level of interest:** A paper of considerable general medical or scientific interest

**Advice on publication:** Accept after discretionary revisions

Compulsory comments

1. **Abstract**
   a. The abstract is clear with clearly stated objective.
   b. The results understandable and believable.
   c. Because calibration was not evaluated, the results reflect only half of what is necessary for the study objectives to be achieved
   d. Conclusions are a fair reflection of the data. However, I suggest for the authors to state that the discrimination is “acceptable to excellent” and add conclusion about the calibration of the prognostic scores after performing calibration analysis. Since calibration is not included, I do not think the authors can conclude that the overall performance of the APACHE systems is superior based on their data.
   e. I believe that the study is interesting, worthwhile, and important since APACHE III has not been well studied in this patient population.

2. **Introduction**
   a. The background has identified the gap in knowledge that warranted the study. The background section is well written giving a nice introduction about the disease severity models for readers who may not be familiar with them. However, the fact that both APACHE II and III were developed and are being used mainly for critically ill patients admitted to the intensive care units should be clearly stated.
   b. I suggest adding a reference or description for the clinical diagnosis of cirrhosis.
   c. The exclusion of patients admitted to the ICU weakens the study because sicker patients likely to die are excluded resulting in mortality rate of only 11.5%. In a study with the primary objective of evaluating the performance of different prognostic models in predicting mortality, the rational of excluding patients with high mortality is not clear to me.
   d. They used APACHE II and III scores, not the predicted mortality rates. The APACHE II and III
prognostic models use not only the raw scores but also the admission diagnoses in the logistic regression models for predicting mortality. Although both patients may have underlying liver disease, the APACHE II and III predicted mortality rate of one patient admitted for peritonitis/sepsis is likely to be different from another patient admitted for GI bleeding. Although the logistic regression coefficients needed for calculating the APACHE III are proprietary and not available, the APACHE II predicted mortality rates could be calculated based on published information in reference 30.

d. In the first paragraph of page 6, the authors have stated: "Numerical values were expressed as mean +/- SD." in line 4 and "Numerical values were expressed as median and ranges." in line 6 and 7. I suggest that they use mean +/- SD for normal and median and ranges for skewed numerical values.
e. Overall, the methods are appropriate for achieving the objective of the study. However, currently, the performance of prognostic models is evaluated by their discrimination and calibration. Discrimination is measured by the AUC and calibration by the Hosmer-Lemeshow statistic. The present study includes results about discrimination but not calibration. Calibration analysis tells us how well a prognostic system discriminates between survivors and non-survivors at various levels (deciles) of risk. If possible, adding calibration analysis with Hosmer-Lemeshow statistic with tables of deciles risk will be more informative. AUC tells only half of the story.
f. Discrimination is classified into different categories based on the AUC. For example, Hosmer and Lemeshow classified AUC between 0.7 and 0.8 as "acceptable" and between 0.8 and 0.9 as "excellent" discrimination in their book (Applied Logistic Regression, second edition, 2000, page 160-164). The present study should use this or similar classification to evaluate the performance of the prognostic models.
g. In the last paragraph of the results section, we are told about the correlation of length of hospital stay and CPS, APACHE II and III score for survivors. The statistical method used to determine the correlation of the scores with the length of hospital stay should be stated in the methods section.

4. Results
a. Only five patients were transferred to ICU during the hospital course. Of the 23 deaths, how many refused or were denied "resuscitation" and ICU admission? How many of the deaths occurred in the ICU?
b. The cause of death stated as "kidney and/or liver" need to be more specific. How many patients died of liver failure alone, kidney failure alone, and the combination of the two?
c. In the results section, the authors repeatedly stated that the CPS, APACHE II and APACHE III "risk stratified" patients with liver cirrhosis. Table 1 shows that there were significant differences in CPS, APACHE II scores, APACHE III scores, and PT between survivors and non-survivors. If the authors want to address the performance of these prognostic systems in discriminating survivors from non-survivors, the answer can be found in Table 2 and figure 1.
d. As I stated above, there are different classifications of discrimination based on the AUC. The authors should use one of those classifications. Using "reliable" as in the last paragraph, page 7, is not adequate.
e. Although the authors defined "statistical significance" as P 0.05 and despite admitting that the differences were not statistically significant, in the last paragraph of page 7, the authors stated that the AUC for CPS was larger than the AUC for APACHE III and the AUC for APACHE III was larger than the AUC for APACHE II. The results should be consistent with the methods.
f. In table 2, I suggest adding 95% before CI and combining the 95% CI and AUC rows as "AUC (95%CI)".

5. Discussion
a. The discussion section can be shortened. Part of the discussion repeats the introduction. The wide 95% CI of the AUC suggests sample size problem, especially when only 23 patients died. Acknowledge this limitation in the discussion.

6. References
a. Delete reference 37 since it is just the abstract form of reference 36.
Discretionary comments
1. Page 3 last paragraph, the authors have stated that APACHE is the "best" severity-of-disease classification. I am not sure it is better than other models such as SAPS.
2. Do not use numbers at the beginning of a sentence.

Competing interests:

None declared.