Meta-Analysis of King's College Criteria and Model for End Stage Liver Disease to Predict Outcome in Acute Liver Failure

Mark JW McPhail^{1,2*}, Hugo Farne^{1*}, Naz Senvar², Julia A Wendon¹ and William Bernal¹

¹Liver Intensive Therapy Unit, Institute of Liver Studies, King's College Hospital, Denmark Hill, London

²Department of Hepatology, St Mary's Hospital, Imperial College London, South Wharf Road, Paddington, London

*These authors contributed equally to this article

Correspondence-

Dr Mark McPhail

Liver Intensive Therapy Unit

Institute of Liver Studies

Kings College Hospital

Denmark Hill

London SE5 9RS

Email m.mcphail@csc.mrc.ac.uk

Phone +44 (0)20 3299 9000

Fax +44 (0)20 3299 3167

Electronic word count - 3154

Number of tables 3 Number of figures 3 (+2 Supplemental)

Abbreviations-

AALF	Acetaminophen Acute Liver Failure
ALF	Acute Liver Failure
DOR	Diagnostic Odds Ratio
ELT	Emergency Liver Transplantation
FPR	False Positive Rate
KCC	King's College Criteria
LR	Likelihood Ratio
MELD	Model for End Stage Liver Disease
NAALF	Non-Acetaminophen Acute Liver Failure
ROC	Receiver Operator Curve
sROC	Summary Receiver Operator Curve
STARD	STAndards for the Reporting of Diagnostic Accuracy
TIPS	Transjugular Intrahepatic Portosystemic Shunt
TPR	True Positive Rate

Disclosures-

MJWM is grateful to the National Institute for Health Research, UK Biomedical Research Centre at Imperial College London for infrastructure support. There are no relevant conflicts of interest.

Grant Support-

MJWM was supported by the Wellcome Trust, UK as part of a Postdoctoral Training Fellowship during the production of this article

Author Contributions-

HF, MM and NS performed the literature search, identified studies for inclusion, extracted the data, and performed the analysis. MM and HF drafted the manuscript. WB ratified the literature search and extracted data. The remaining authors revised the manuscript. WB is the guarantor.

ABSTRACT

Background & Aims

Prognostication in Acute Liver Failure (ALF) aims to identify patients who require Emergency Liver Transplantation (ELT). We assessed the accuracy of King's College Criteria (KCC) versus the Model-for-End-Stage-Liver-Disease (MELD) in ALF through meta-analysis of studies which report the accuracy of both tests.

Methods

After systematic literature search and quality assessment collated data was meta-analysed for pooled sensitivity, specificity, Diagnostic Odds Ratio (DOR) and summary Receiver Operator Curve (sROC) analysis. Subgroup analysis was defined by study quality; era; size; aetiology (acetaminophen v non-acetaminophen AALF v NAALF); location (Europe vs. non-Europe), ELT centre, and inclusion of patients undergoing ELT.

Results

Twenty three studies with data for 2153 patients published between 2006 and 2015 were retrieved. The DOR for KCC was 5.3 (95% CI 3.7-7.6, heterogeneity 57%) and 7.0 (5.1-9.7, heterogeneity 48%) for MELD indicating comparable overall accuracy; the sAUROC for KCC was 0.76 and 0.78 for MELD. For AALF the sensitivity was 58(51-65)% and specificity 89(85-93)% for KCC, vs. sensitivity 80 (74-86)% and specificity 53(47-59)% for MELD. For NAALF the sensitivity was 58(54-63) % and specificity 74(69-78)% for KCC, vs. sensitivity 76(72-80)% and specificity 73 (69-78)% for MELD. The DOR for KCC in cases of AALF was 10.4(4.9-22.1) and for MELD 6.6 (2.1-20.2) whereas for NAALF the DOR for KCC was 4.16 (2.34-7.40) and 8.42 (5.98-11.88) for MELD.

Conclusions

Although KCC performs better for AALF, MELD has improved prognostic accuracy in NAALF with significant heterogeneity in published studies and neither is optimal in all scenarios. Given the critical importance of loss of specificity in ELT wait-listing decisions, MELD should not replace KCC in AALF, but may have a role in NAALF.

Keywords Acute liver failure; Meta-analysis; Liver transplantation.

INTRODUCTION

Acute Liver Failure (ALF) is a rare, but devastating illness with a high risk of progression to multi-organ failure and death¹⁻³. Outcomes in ALF improved with the advent of Emergency Liver Transplantation (ELT) in the 1980s⁴, and further improvements have occurred recently with modern intensive care practices⁵. The key clinical issue remains to accurately identify patients with ALF who will die without ELT, and those who will survive with medical management alone. Failure to identify those in need of ELT results in a potentially preventable death. Conversely, wrongly classifying prospective survivors as in need of ELT subjects the patient to the morbidity and mortality associated with transplantation. In addition a graft is removed from the donor pool.

Multiple prognostic systems have been developed to aid decision making for ELT. The longest established are the King's College Criteria (KCC)⁶. These were derived from a retrospective analysis of 588 patients from the pre-transplant era 1973-1985, and have subsequently been validated in many centres⁷⁻¹¹. Recognising that there are differences in the natural history of Acetaminophen-induced ALF (AALF) and Non-Acetaminophen-induced ALF (NAALF), separate criteria were derived for each. The KCC demonstrate high specificity, but have been criticised for their low sensitivity and negative predictive value for a poor outcome^{12, 13}. Nevertheless these easy-to-use criteria have consistently outperformed novel scoring systems and remain the benchmark for prognostication in ALF.

With concerns about the low sensitivity of the KCC to predict mortality in ALF, clinicians have sought alternatives that might outperform the KCC, including the Model for End-stage Liver Disease (MELD) score¹⁴⁻¹⁶. Early reports noted the moderate accuracy of MELD as an alternative to the KCC in ALF¹⁷. However it appears that whilst MELD has higher sensitivity for predicting death in ALF than the KCC, it may have a lower specificity for predicting survival¹⁸.

While a body of evidence has emerged comparing the KCC and MELD, controversy remains over which scoring system has the superior diagnostic accuracy^{19, 20}. The conflicting results to date may be explained by differences in design and relatively small sample sizes (the largest study had 380 patients). One particularly salient difference is the treatment of transplanted patients. Authors must decide whether to include patients with ALF who received ELT as non-survivors, which may not necessarily be the case, or to exclude them altogether – in either case introducing bias which is not present in reports from the pre-ELT era. There are also differences in study location (with differing resources and guidelines in each, and a lack of international standardisation of laboratory variables^{21, 22}), the date range reviewed (with possible changes in disease spectrum and efficacy of supportive management over time), and aetiologies included (all ALF or a subset i.e. AALF, NAALF, viral hepatitis, or anti-tuberculous chemotherapy-induced ALF).

To date, there have been three meta-analyses²³ of the performance of the KCC in ALF. The first included only Acetaminophen-induced ALF (AALF) identifying nine studies in total, and concluded that the KCC had limited sensitivity¹³. A later study updated this meta-analysis for AALF to include a further five studies, and found that the KCC had high diagnostic accuracy²⁴. However given previous reports have found the KCC to be superior in AALF than in the subset with Non-Acetaminophen-induced ALF (NAALF),²⁵ a later meta-analysis assessed the KCC in NAALF, again finding that the KCC have high overall accuracy and in particular good specificity if more limited sensitivity²⁶.

There are no published meta-analyses assessing the performance of MELD in ALF, nor any comparing MELD and the KCC. The aim of this study was to quantitatively assess and compare the prognostic accuracy of KCC and MELD in ALF through a meta-analysis of all studies reporting the diagnostic accuracy of both prognostic schema in the same population.

METHODS

Search methodology

We conducted a systematic review of the literature by searching MEDLINE via PUBMED, Scopus, Web of Science, CINAHL and EMBASE for articles that reported on the diagnostic accuracy of KCC and MELD for ALF published between 1 January 2001 (as 2001 was the year that MELD was first validated as a predictor of outcome in chronic liver disease) and 1 May 2015. The search terms used were “Acute liver failure” OR “Fulminant liver failure”; “King’s Criteria” OR “King’s College Criteria” OR “Kings College Criteria” OR “King’s College Hospital Criteria”; “Model for End-stage Liver Disease” OR “MELD”; AND (“Outcome” OR “Prediction” OR “Sensitivity” OR “Specificity” OR “Diagnostic accuracy”). All adult studies in humans were included. In addition, we obtained potentially relevant articles by tracking references from the original articles and review articles. Only data from the primary article were used in the meta-analysis. All potential articles were assessed independently by two researchers (HF, MM) according to prospectively defined eligibility criteria, and disagreements were resolved by consensus or consultation with a third author (WB).

Data extraction

We used a pre-defined form for recording details of study design, data quality, and test performance. The following study design details were extracted for each study: country, sample size, timeframe, aetiologies of ALF, inclusion of ELT as an outcome equivalent to death or not, and the MELD cut-off score used. Information on data quality were extracted as per the STAndards for the Reporting of Diagnostic Accuracy Studies (STARD) pro forma²⁷. Test performance data were extracted as a 2 x 2 table of true negative, true positive, false negative and false positive, if possible directly from the tabulated results. If these were not available, they were calculated from marginal totals (e.g. total number of survivors) and reported sensitivity, specificity, positive predictive value, and/or negative predictive values. If this was not possible or there was doubt over the 2 x 2 calculation, the study was

excluded from the subsequent analysis. Original authors were contacted over doubtful data extraction. Discrepancies were resolved by consensus of two separate reviewers (HF and MM).

Assessment of study quality

Retrieved studies were quality assessed using a modified STAndards for the Reporting of Diagnostic Accuracy Studies (STARD) pro forma (See Supplementary Material). We modified this scoring system from 25 to 22 as three points were not relevant to this study (4, 17 and 20). Two readers (HF and MM) independently assessed all included studies according to the pre-defined pro forma. Studies were then segregated into 'low' or 'high' methodological quality depending on whether they met greater than 50% of the study quality pro forma.

Data analysis

The DerSimonian-Laird random effects method was used to produce summary estimates of sensitivity, specificity, likelihood ratios (LR) and diagnostic odds ratio (DOR, defined as the ratio of positive to negative likelihood ratios). The value of a DOR ranges from zero to infinity, with higher values indicating better discriminatory test performance. Where required, a zero cell correction of 0.5 was added to all cells to prevent computational problems arising where denominators were equal to zero²⁸. Confidence intervals for sensitivity and specificity were calculated using F-distribution method for the binomial proportion²⁹. For the DOR confidence intervals were calculated assuming normality of the distribution of the natural logarithm of the DOR. The summary receiver operator curve (sROC) was used to graphically determine performance following testing for correlation between sensitivity and specificity (as logit true positive rate (TPR) vs. logit false positive rate (FPR)) to explore for threshold effects and subsequent assessment for constant DOR using the Moses-Shapiro-Littenburg model³⁰. Where the model suggested the DOR varied with the diagnostic threshold, an asymmetrical sROC curve was fitted (instead of a symmetrical curve).

Heterogeneity was investigated using pre-planned subgroup analysis and the I^2 method for calculating residual heterogeneity independent of study sample size and chance³¹. Pre-planned subgroups were defined according to: study quality (low vs. high STARD score (50% criteria met as cut-off)), inclusion of ELT as an outcome equal to death or not; study era (early or late, defined as whether the median recruitment year was pre- or post-2002) as a proxy for improvements in critical care; study size (n=0-50, 50-100 or >100); study location (Europe or non-European); and the aetiology of the ALF (AALF vs. NAALF).

A funnel plot and effective sample size (ESS) regression analysis (the logarithm of the DOR plotted against $1/\sqrt{\text{ESS}}$) was used to investigate publication bias. $1/\sqrt{\text{ESS}}$ is proportional to root $(1/n_1 + 1/n_2)$ where n_1 = number diseased died) and n_2 = number not diseased (survived)³². Data analyses were performed using the freeware Meta-Disc version 1.4 (Universidad Complutense, Madrid, Spain) and Eggers statistic calculated in Excel (Microsoft Corporation, Redmond WA)³³.

RESULTS

The search strategy identified 4,063 potentially relevant studies. After screening their title and abstract, 311 were retrieved and evaluated. A further nine studies were identified from the references of the articles evaluated and also reviewed. 23 studies were found with enough information to go forward into a meta-analysis (see Figure 1)^{19, 20, 34-54}. The eligible studies evaluated a total of 2,153 patients. KCC was calculated for all patients, but MELD was only available for 2,101 patients because two studies had insufficient data to calculate MELD for every patient^{38, 45}. From the data presented it was impossible to define patient level KCC and MELD scores thus it was decided to include overall totals.

The characteristics of studies included in the meta-analysis are shown in Table 1 with further details in supplementary table 2. 10/23 studies were deemed to be of high quality according to the modified STARD pro forma (STARD score >12). 10/23 included ELT as a composite endpoint with death. 11/23 were from an early era, with the median year of admission of patients included pre-2002. 5/23 were small (n<50), 10/23 medium (n=50-100), and 8/23 large studies (n>100). 10/23 were conducted in Europe. 5/23 studies included both AALF and NAALF without presenting data for each subgroup. Of the remaining studies, four were limited to AALF, seven to NAALF, and two included both but presented the data for AALF and NAALF separately. MELD thresholds were not standardised across the studies, ranging from 25 to 37. In 8 studies transplantation was not performed at that centre. All other studies were reports from transplant centres. Nine studies had a prospective design with the remainder retrospective observational cohort studies.

Sensitivity, specificity and diagnostic odds ratio

Table 2 and Figure 2a-d show the results of individual and pooled sensitivity and specificity for KCC and MELD from all data gathered, ordered by year of publication. The pooled sensitivity for all 23

studies for KCC was 59 (95% CI 56-62) % and specificity 79 (77-81)%, vs. sensitivity 74 (71-77)% and specificity 67 (64-69)% for MELD. The pooled positive LR for KCC was 2.5 (2.0-3.2) and negative LR 0.54 (0.46-0.64), vs. positive LR 2.4 (2.0-3.0) and negative LR 0.38 (0.31-0.46) for MELD. The pooled DOR for KCC was 5.3 (3.7-7.7) and for MELD was 7.0 (5.1-9.7).

The Spearman correlation for logit (TPR) vs logit (FPR) for KCC was 0.363 ($p=0.082$) and for MELD was 0.193 ($p=0.378$) indicating that within each method of prediction there was little evidence of explicit threshold effects. Across all of the studies the correlation was 0.439, $p=0.007$ reflecting the different thresholds between the two methods. Using the Moses-Shapiro-Litenberg method to assess for constant DOR gave a p value of 0.689 for KCC and 0.264 for MELD indicating that the DOR was constant. Again across all studies this p value was significant 0.034.

An asymmetrical summary Receiver Operating Curve (ROC) was therefore the most appropriate representation of global diagnostic accuracy (Figure 3). The summary Area Under the ROC (AUROC) was 0.76 for KCC and 0.78 for MELD, with a Q statistic (SE) of 0.024 and 0.018 respectively.

Heterogeneity analysis & subgroup analysis

The heterogeneity in the sensitivity, specificity and DOR was broadly equivalent for KCC and MELD (for sensitivity KCC 84% vs. MELD 76%; for specificity KCC 76% vs. MELD 72%; for DOR KCC 59% vs. MELD 48%, see Figure 4 (Supplementals)).

Subgroup analysis was performed to assess differences in heterogeneity and diagnostic accuracy between the groups specified earlier. Tables 3 (A and B) shows the sensitivity, specificity, positive and negative likelihood ratios, and diagnostic odds ratio estimates for the subgroups considered. Of note, the specificity for MELD was further reduced in patients with AALF (53(47-59)% vs 73(69-78)% for NAALF with the DOR for the two schema in AALF patients being 10.4 (4.9-22.1) for KCC and 6.6 (2.1-

20.2) for MELD. The heterogeneity for MELD was reduced in later studies, those of higher quality and studies performed outside Europe. Lower quality studies tended to give reduce DOR for both KCC and MELD. The specificity for studies from ELT centres was lower for KCC than those centres where ELT was not available (83 (80-86)% vs 73(69-77)% although this pattern was not reproduced for MELD.

Publication bias

Examination of the Deeks' funnel plots for KCC and MELD (Figure 5 (Supplementals)) revealed no evidence of publication bias. The funnel plots show that when the precision of the study (stated as a function of sample size) is plotted against diagnostic test accuracy (as a function of the DOR), there is a symmetrical spread of studies. In particular, there are no gaps in the lower right-hand side to suggest that smaller studies not favouring the test (KCC or MELD respectively) are missing. Furthermore Egger's statistic was not significant again suggesting publication bias was not present.

DISCUSSION

This meta-analysis confirms that when comparing KCC and MELD for outcome prediction in ALF KCC have lower sensitivity and MELD lower specificity. However, the overall diagnostic accuracy of KCC and MELD in determining outcome as measured by DOR and sROC is comparable. The particular strength of the KCC appears to be in AALF where the DOR is twice that of MELD.

There is significant heterogeneity in the published studies for the performance of both prediction schemes but particularly for MELD, which may be related to differing thresholds for poor outcome. Indeed the use of different thresholds makes the pooling of sensitivities, specificities, likelihood and diagnostic ratios potentially problematic³⁰. While threshold analysis confirmed the difference between KCC and MELD but did not suggest an explicit threshold problem with MELD, we proceeded as if an implicit threshold difference was occurring to mitigate this potential source of bias. As these measures change in an interdependent manner with variations in the diagnostic threshold, plotting the results at different thresholds allows a curve to be fitted that fully encapsulates this (an sROC curve). The sROC analysis is therefore a more valid way to pool the results of studies with varying thresholds. In our analysis, this suggested the KCC had a small superiority to MELD (AUROC 0.78 vs. 0.76 respectively). The subgroup analysis are therefore of great importance in delineating the different prognostic properties of KCC and MELD as they explore sources of bias albeit at the cost of statistical power through reduced numbers of patients.

There are important differences in the application of the KCC and MELD in AALF and NAALF. As previously observed²⁵, the KCC show greater diagnostic accuracy (higher DOR) in AALF than NAALF (10.87 vs. 4.89), consistent with the higher specificity in this group (89% vs. 78%). The same is not true of MELD; the DOR is lower in AALF (5.78) than NAALF (7.84). The explanation may lie in the natural history of AALF, in which the acuity of the insult and early presentation results in a more hepatotoxic

early phase reflected in significant transaminitis with a delayed rise in bilirubin. Indeed, bilirubin at presentation has not equivocally been shown to be prognostic^{6, 55}. This is reflected in the KCC, which do not include bilirubin as a prognostic marker in AALF, whereas MELD has bilirubin as a central component. This is no doubt a consequence of the fact that the KCC were derived from an ALF cohort, whereas MELD was developed from results in chronic liver disease patients undergoing TIPS. Nevertheless while non AALF cases include sub-acute liver failure where jaundice is a key marker of poor prognosis, MELD does not seem to have a clear advantage over KCC in NAALF cases in terms of the DOR. This is likely because of the specific non AALF KCC which have encapsulated this pathophysiological difference explicitly. Recently changes to Clichy criteria have also been suggested in a similar manner⁵⁶. We also noted that while some bias exists in centres where ELT is available for the specificity of KCC, other measures of KCC performance and data for MELD were not different. There may therefore be a case for using MELD preferentially in cases of NAALF given the higher DOR seen in this meta-analysis.

Strategically, using MELD for NAALF and KCC for AALF would also require considerable thought to the availability of organs in particular countries⁵⁶. Where supply is short a specific score like KCC may be preferred and therefore it would be unlikely that an organ would be utilised in a patient who would otherwise have survived (low false positive rate). Where organ shortages are not so acute then the opposite would be case and a highly sensitive score may be preferred (low false negative rate) and the chance of missing the opportunity to transplant a patient would be minimised. Ideally no such strategic conflict would exist but as a) neither MELD or KCC are perfect scores and b) organ shortages do occur in some countries organ allocation policies would need to take these factors and the aetiologies prevalent within each area. This may be why KCC is preferred in countries facing such organ shortages and with high rates of AALF. We hope these data help inform such decisions and future research.

It has been argued that the inclusion of a subjective measure (encephalopathy) in the KCC weakens their usefulness compared to MELD, which is based entirely on objective laboratory values for bilirubin, creatinine and INR^{36, 40}. In theory this means the KCC can be modulated by the clinician's interpretation of West Haven criteria. However the KCC can be calculated almost immediately using clinical criteria and bedside tests such as arterial blood gases. A worsening grade of HE can be detected at the bedside and incorporated into KCC without awaiting further biochemical analysis. Clearly such delays are relatively short but in cases of fulminant hepatic failure it is clearly advantageous to use simpler bedside tests during the evolution of disease. While modifications to MELD exist where HE is incorporated these modifications to MELD are not used in listing decisions in ALF. Not incorporating this key prognostic evolution may both reduce predictive accuracy and delay the decision to list for OLT.

There remain a number of limitations in this analysis. Information on prothrombin time measurements and assay details were not available in all studies and may have contributed to heterogeneity or threshold effects. Not all studies had exact parity in the number of patients where KCC or MELD were available. The number missing was 52 patients was less than 3% of the total cohort so was not likely to significantly impact the overall conclusions. Although we do perform analysis to explore threshold effects the use of different MELD cut-offs does suggest caution is necessary in interpreting these results. Data on each patient would be required to both re-explore MELD cut-off values over these cohorts and indeed make a combined model of MELD and KCC.

There are no published studies to date looking at the use of the KCC and MELD in combination. The potential benefits of combining the specificity of the KCC with the sensitivity of MELD are attractive. But it is unclear how this would work in practice: in those instances where the scores were incongruous, which would assume primacy? Such a study could have potential uses in cases where clinicians have a strong suspicion of a poor outcome but are unable to seek listing due to existing

waitlist registration criteria. It was not possible to assess the combination of KCC and MELD from the published studies. Such novel methods would require data for each patient rather than summative as presented for publication. Further studies on the use of MELD in cases with negative KCC or vice versa would be of interest to the transplant community and could be an area of future investigation.

What is clear is that neither KCC nor MELD are the perfect prognostic tool in ALF. While their clinical or biochemical simplicity is useful in using these schema, their ability to accurately capture biological responses is part of their limitations. Many new biomarkers have been proposed in ALF but have failed to be validated in larger studies or are deemed not ready for widespread distribution. Also the original derivation of the KCC remains one of the largest studies ever performed in this field and coupled with the strength of statistical analysis underpins its ongoing validity. Future studies to generate biological markers of similar validity should focus on collaborative work and biobanking to yield useful tools in larger cohorts with high translational potential.

In conclusion, following this first quantitative comparison of outcome prediction in ALF comparing KCC and MELD we suggest that MELD should not replace KCC in AALF but may give improved prognostic accuracy in NAALF. Neither KCC nor MELD are optimal in all circumstances so there remains an urgent need for more accurate outcome prediction systems in ALF.

Figure Legends

Figure 1: Flow diagram of assessment of studies identified in the systematic review

Figure 2: Pooled sensitivity for a) Kings College Criteria (KCC) and b) Model for End Stage Liver Disease (MELD). Pooled specificity for c) Kings College Criteria (KCC) and d) Model for End Stage Liver Disease (MELD)

Figure 3: Asymmetric summary Receiver Operator Curve (sROC) for a) Kings College Criteria (KCC) and b) Model for End Stage Liver Disease (MELD)

Figure 4 (Supplementals): Pooled diagnostic odds ratio (DOR) for for a) Kings College Criteria (KCC) and b) Model for End Stage Liver Disease (MELD)

Figure 5 (Supplementals): Deeks' Funnel Plot for a) Kings College Criteria (KCC) and b) Model for End Stage Liver Disease (MELD)

REFERENCES

1. Bernal W, Auzinger G, Dhawan A, et al. Acute liver failure. *Lancet*. 2010;376(9736):190-201.
2. Lee WM. Acute liver failure. *The New England journal of medicine*. 1993;329(25):1862-72.
3. Ostapowicz G, Fontana RJ, Schiodt FV, et al. Results of a prospective study of acute liver failure at 17 tertiary care centers in the United States. *Ann Intern Med*. 2002;137(12):947-54.
4. O'Grady J. Liver transplantation for acute liver failure. *Best practice & research Clinical gastroenterology*. 2012;26(1):27-33.
5. Bernal W, Auzinger G, Sizer E, et al. Intensive care management of acute liver failure. *Semin Liver Dis*. 2008;28(2):188-200.
6. O'Grady JG, Alexander GJ, Hayllar KM, et al. Early indicators of prognosis in fulminant hepatic failure. *Gastroenterology*. 1989;97(2):439-45.
7. Riordan SM, Williams R. Use and validation of selection criteria for liver transplantation in acute liver failure. *Liver transplantation : official publication of the American Association for the Study of Liver Diseases and the International Liver Transplantation Society*. 2000;6(2):170-3.
8. Shakil AO, Kramer D, Mazariegos GV, et al. Acute liver failure: clinical features, outcome analysis, and applicability of prognostic criteria. *Liver transplantation : official publication of the American Association for the Study of Liver Diseases and the International Liver Transplantation Society*. 2000;6(2):163-9.
9. Donaldson BW, Gopinath R, Wanless IR, et al. The role of transjugular liver biopsy in fulminant liver failure: relation to other prognostic indicators. *Hepatology*. 1993;18(6):1370-6.
10. Pauwels A, Mostefa-Kara N, Florent C, et al. Emergency liver transplantation for acute liver failure. Evaluation of London and Clichy criteria. *J Hepatol*. 1993;17(1):124-7.
11. Anand AC, Nightingale P, Neuberger JM. Early indicators of prognosis in fulminant hepatic failure: an assessment of the King's criteria. *J Hepatol*. 1997;26(1):62-8.
12. Lee WM. Acute liver failure in the United States. *Semin Liver Dis*. 2003;23(3):217-26.
13. Bailey B, Amre DK, Gaudreault P. Fulminant hepatic failure secondary to acetaminophen poisoning: a systematic review and meta-analysis of prognostic criteria determining the need for liver transplantation. *Critical care medicine*. 2003;31(1):299-305.
14. Kamath PS, Wiesner RH, Malinchoc M, et al. A model to predict survival in patients with end-stage liver disease. *Hepatology*. 2001;33(2):464-70.
15. Malinchoc M, Kamath PS, Gordon FD, et al. A model to predict poor survival in patients undergoing transjugular intrahepatic portosystemic shunts. *Hepatology*. 2000;31(4):864-71.
16. Wiesner R, Edwards E, Freeman R, et al. Model for end-stage liver disease (MELD) and allocation of donor livers. *Gastroenterology*. 2003;124(1):91-6.
17. Kremers WK, van IM, Kim WR, et al. MELD score as a predictor of pretransplant and posttransplant survival in OPTN/UNOS status 1 patients. *Hepatology*. 2004;39(3):764-9.
18. Rossaro L CC, Polson J, Bowlus CL, Hynan LS, Fontana RJ, Shakil AO, Stravitz T, Schilsky ML, Hay EJ, Lee WM, and the ALF Study Group. Performance of MELD in Predicting Outcome in Acute Liver Failure. *Gastroenterology*. 2005;128(4):A705.
19. Taylor RM, Davern T, Munoz S, et al. Fulminant hepatitis A virus infection in the United States: Incidence, prognosis, and outcomes. *Hepatology*. 2006;44(6):1589-97.
20. Kumar R, Shalimar, Sharma H, et al. Prospective derivation and validation of early dynamic model for predicting outcome in patients with acute liver failure. *Gut*. 2012;61(7):1068-75.
21. Newsome PN, Henderson NC, Germain L, et al. Prothrombin time to assess fulminant hepatic failure. *Lancet*. 2001;358(9299):2172.
22. Ravaioli M, Masetti M, Ridolfi L, et al. Laboratory test variability and model for end-stage liver disease score calculation: effect on liver allocation and proposal for adjustment. *Transplantation*. 2007;83(7):919-24.
23. Irwig L, Tosteson AN, Gatsonis C, et al. Guidelines for meta-analyses evaluating diagnostic tests. *Ann Intern Med*. 1994;120(8):667-76.

24. Craig DG, Ford AC, Hayes PC, et al. Systematic review: prognostic tests of paracetamol-induced acute liver failure. *Aliment Pharmacol Ther.* 2010;31(10):1064-76.
25. Polson J, Lee WM, American Association for the Study of Liver D. AASLD position paper: the management of acute liver failure. *Hepatology.* 2005;41(5):1179-97.
26. McPhail MJ, Wendon JA, Bernal W. Meta-analysis of performance of Kings's College Hospital Criteria in prediction of outcome in non-paracetamol-induced acute liver failure. *J Hepatol.* 2010;53(3):492-9.
27. Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Family practice.* 2004;21(1):4-10.
28. Dinnes J, Deeks J, Kirby J, et al. A methodological review of how heterogeneity has been examined in systematic reviews of diagnostic test accuracy. *Health technology assessment.* 2005;9(12):1-113, iii.
29. Leemis LM TK. A comparison of approximate interval estimators for the Bernoulli parameter. *The American Statistician.* 1994;50:63-8.
30. Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. *Statistics in medicine.* 1993;12(14):1293-316.
31. Higgins JP, Thompson SG, Deeks JJ, et al. Measuring inconsistency in meta-analyses. *Bmj.* 2003;327(7414):557-60.
32. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *Journal of clinical epidemiology.* 2005;58(9):882-93.
33. Zamora J, Abraira V, Muriel A, et al. Meta-DiSc: a software for meta-analysis of test accuracy data. *BMC medical research methodology.* 2006;6:31.
34. Pelaez-Luna M, Martinez-Salgado J, Olivera-Martinez MA. Utility of the MAYO End-Stage Liver Disease score, King's College Criteria, and a new in-hospital mortality score in the prognosis of in-hospital mortality in acute liver failure. *Transplantation proceedings.* 2006;38(3):927-9.
35. Zaman MB, Hoti E, Qasim A, et al. MELD score as a prognostic model for listing acute liver failure patients for liver transplantation. *Transplantation proceedings.* 2006;38(7):2097-8.
36. Dhiman RK, Jain S, Maheshwari U, et al. Early indicators of prognosis in fulminant hepatic failure: an assessment of the Model for End-Stage Liver Disease (MELD) and King's College Hospital criteria. *Liver transplantation : official publication of the American Association for the Study of Liver Diseases and the International Liver Transplantation Society.* 2007;13(6):814-21.
37. Katoonizadeh A, Decaestecker J, Wilmer A, et al. MELD score to predict outcome in adult patients with non-acetaminophen-induced acute liver failure. *Liver Int.* 2007;27(3):329-34.
38. Schmidt LE, Larsen FS. MELD score as a predictor of liver failure and death in patients with acetaminophen-induced liver injury. *Hepatology.* 2007;45(3):789-96.
39. Wei G, Bergquist A, Broome U, et al. Acute liver failure in Sweden: etiology and outcome. *J Intern Med.* 2007;262(3):393-401.
40. Yantorno SE, Kremers WK, Ruf AE, et al. MELD is superior to King's college and Clichy's criteria to assess prognosis in fulminant hepatic failure. *Liver transplantation : official publication of the American Association for the Study of Liver Diseases and the International Liver Transplantation Society.* 2007;13(6):822-8.
41. Hadem J, Stiefel P, Bahr MJ, et al. Prognostic implications of lactate, bilirubin, and etiology in German patients with acute liver failure. *Clin Gastroenterol Hepatol.* 2008;6(3):339-45.
42. Yamagishi Y, Saito H, Ebinuma H, et al. A new prognostic formula for adult acute liver failure using computer tomography-derived hepatic volumetric analysis. *Journal of gastroenterology.* 2009;44(6):615-23.
43. Bechmann LP, Jochum C, Kocabayoglu P, et al. Cytokeratin 18-based modification of the MELD score improves prediction of spontaneous survival after acute liver injury. *J Hepatol.* 2010;53(4):639-47.

44. Chen X-B ZX, Tang H. Value of KCH criteria, MELD score and MELDNa score for prediction of prognosis of acute liver failure. *World Chinese Journal of Digestology*. 2011;19(8):855-9.
45. Kumar R, Shalimar, Bhatia V, et al. Antituberculosis therapy-induced acute liver failure: magnitude, profile, prognosis, and predictors of outcome. *Hepatology*. 2010;51(5):1665-74.
46. Cholongitas E, Theocharidou E, Vasianopoulou P, et al. Comparison of the sequential organ failure assessment score with the King's College Hospital criteria and the model for end-stage liver disease score for the prognosis of acetaminophen-induced acute liver failure. *Liver transplantation : official publication of the American Association for the Study of Liver Diseases and the International Liver Transplantation Society*. 2012;18(4):405-12.
47. Craig DG, Reid TW, Wright EC, et al. The sequential organ failure assessment (SOFA) score is prognostically superior to the model for end-stage liver disease (MELD) and MELD variants following paracetamol (acetaminophen) overdose. *Aliment Pharmacol Ther*. 2012;35(6):705-13.
48. Hadem J, Tacke F, Bruns T, et al. Etiologies and outcomes of acute liver failure in Germany. *Clin Gastroenterol Hepatol*. 2012;10(6):664-9 e2.
49. Parkash O, Mumtaz K, Hamid S, et al. MELD score: utility and comparison with King's College criteria in non-acetaminophen acute liver failure. *Journal of the College of Physicians and Surgeons--Pakistan : JCPSP*. 2012;22(8):492-6.
50. Manka P, Bechmann LP, Tacke F, et al. Serum sodium based modification of the MELD does not improve prediction of outcome in acute liver failure. *BMC gastroenterology*. 2013;13:58.
51. Mendizabal M, Marciano S, Videla MG, et al. Changing etiologies and outcomes of acute liver failure: perspectives from 6 transplant centers in Argentina. *Liver transplantation : official publication of the American Association for the Study of Liver Diseases and the International Liver Transplantation Society*. 2014;20(4):483-9.
52. Feng HL, Li Q, Wang L, et al. Indocyanine green clearance test combined with MELD score in predicting the short-term prognosis of patients with acute liver failure. *Hepatobiliary & pancreatic diseases international : HBPDI*. 2014;13(3):271-5.
53. Bagchi A, Kumar S, Ray PC, et al. Predictive value of serum actin-free Gc-globulin for complications and outcome in acute liver failure. *Journal of viral hepatitis*. 2015;22(2):192-200.
54. Shaikh S, Quazi, I, Baloch, GH. Comparison of King's College Hospital Criteria (KCH) with Model for End-Stage Liver Disease (MELD) for Predicting Outcome in Patients with Acute Liver Failure *J Med Sci*. 2012;28(4):700-5.
55. Christensen E, Bremmelgaard A, Bahnsen M, et al. Prediction of fatality in fulminant hepatic failure. *Scandinavian journal of gastroenterology*. 1984;19(1):90-6.
56. Ichai P, Legeai C, Francoz C, et al. Patients with acute liver failure listed for superurgent liver transplantation in France: reevaluation of the Clichy-Villejuif criteria. *Liver transplantation : official publication of the American Association for the Study of Liver Diseases and the International Liver Transplantation Society*. 2015;21(4):512-23.

Figure 1: Flow diagram of assessment of studies identified in the systematic review

Figure 2: Pooled sensitivity for a) Kings College Criteria (KCC) and b) Model for End Stage Liver Disease (MELD). Pooled specificity for c) Kings College Criteria (KCC) and d) Model for End Stage Liver Disease (MELD)

Figure 3: Asymmetric summary Receiver Operator Curve (sROC) for a) Kings College Criteria (KCC) and b) Model for End Stage Liver Disease (MELD)

Study	Size	Quality	OLT	OLT centre	*Study era	Country	Aetiology
	(n)	(High = STARD >12)	excluded?		(Early = pre-2002)		
Pelaez-Luna et al 2006	58	Low	Yes	No	Early	Mexico	All
Taylor et al 2006	29	High	No	Yes	Early	US	Hepatitis A
Zaman et al 2006	60	Low	No	Yes	Early	Ireland	POD
Dhiman et al 2007	144	High	Yes	No	Early	India	Viral hepatitis
Katoonizadeh et al 2007	53	Low	Yes	Yes	Early	Belgium	Non-POD
Schmidt & Larsen 2007	124	High	No	Yes	Early	Denmark	POD
Wei et al 2007	229	Low	Yes	Yes	Early	Sweden	All
Yantorno et al 2007	22	Low	Yes	Yes	Early	Argentina	Non-POD
Hadem et al 2008	102	High	No	Yes	Early	Germany	All
Yamagishi et al 2009	30	Low	No	Yes	Late	Japan	Non-POD
Bechmann et al 2010	68	High	No	Yes	Late	Germany	All
Kumar et al 2010	70	Low	Yes	No	Early	India	Anti-TB therapy
Chen et al 2011	37	Low	Yes	No	Late	China	All
Cholongitas et al 2012	125	High	No	Yes	Early	Greece	POD
Craig et al 2012	66	High	No	Yes	Late	UK	POD
Hadem et al 2012	109	Low	No	Yes	Late	Germany	All
Kumar et al 2012	380	High	Yes	No	Late	India	All
Parkash et al 2012	91	Low	Yes	No	Late	Pakistan	Non-POD

Shaikh et al 2012	76	Low	Yes	No	Late	Pakistan	Non-POD
Manka et al 2013	108	High	No	Yes	Late	Germany	All
Feng et al 2014	69	Low	Yes	No	Late	China	Non-POD
Mendizabal et al 2014	71	High	Yes	Yes	Late	Argentina	Non-POD
Bagchi et al 2015	32	Low	Yes	No	Late	India	Non-POD

Table 1 Characteristics of studies included in the meta-analysis (STARD - standards for reporting of diagnostic accuracy studies, AALF- acetaminophen related acute liver failure, NAALF (non-acetaminophen related acute liver failure, TB – tuberculosis, *midway year of study recruitment).

Study	KCC			MELD		
	Sensitivity	Specificity	DOR	Sensitivity	Specificity	DOR
Pelaez-Luna et al 2006	0.46	0.70	2.02	0.71	0.73	6.88
Taylor et al 2006	0.31	0.94	6.67	0.54	0.88	8.17
Zaman et al 2006	0.73	1.00	-	0.97	0.66	50.91
Dhiman et al 2007	0.47	0.88	6.73	0.76	0.67	6.55
Katoonizadeh et al 2007	0.87	0.64	12.92	0.86	0.75	21.78
Schmidt & Larsen 2007	0.55	0.83	5.76	0.58	0.54	1.63
Wei et al 2007	0.30	0.90	3.80	0.70	0.68	4.83
Yantorno et al 2007	0.73	0.82	12.00	0.91	0.82	45.00
Hadem et al 2008	0.58	0.82	6.51	0.65	0.69	4.19
Yamagishi et al 2009	0.82	0.62	7.47	0.41	0.77	2.33
Bechmann et al 2010	0.22	0.68	0.61	0.77	0.76	11.08
Kumar et al 2010	0.34	0.74	1.46	0.73	0.67	5.50
Chen et al 2011	0.65	0.64	3.31	0.85	0.82	24.75
Cholongitas et al 2012	0.47	0.87	5.61	0.89	0.25	2.95
Craig et al 2012	0.86	0.82	26.57	1.00	0.50	-
Hadem et al 2012	0.44	0.76	2.50	0.51	0.79	3.86
Kumar et al 2012	0.62	0.73	4.34	0.73	0.66	5.34
Parkash et al 2012	0.88	0.71	17.14	0.79	0.71	9.00
Shaikh et al 2012	0.85	0.24	1.71	0.78	0.82	16.51

Manka et al 2013	0.72	0.90	24.27	0.78	0.76	10.97
Feng et al 2014	0.49	0.83	4.71	0.70	0.81	9.53
Mendizabal et al 2014	0.73	0.63	4.57	0.85	0.66	10.77
Bagchi et al 2015	0.60	0.86	9.00	0.92	0.86	69.00
Pooled results (95% CI)	0.59 (0.56-0.62)	0.79 (0.77-0.81)	5.27 (3.67-7.57)	0.74 (0.71-0.77)	0.67 (0.64-0.69)	7.01 (5.06-9.71)

Table 2 Individual and pooled sensitivity, specificity, and diagnostic odds ratios (DOR) for King's College Criteria (KCC) and Model for End Stage Liver Disease (MELD).

	No. of studies	Sensitivity (95% CI)	Specificity (95% CI)	Positive LR (95% CI)	Negative LR (95% CI)	DOR (95% CI)	I ² (%)
A) KCC							
	No. of studies	Sensitivity (95% CI)	Specificity (95% CI)	Positive LR (95% CI)	Negative LR (95% CI)	DOR (95% CI)	I ² (%)
All	23	0.59 (0.56-0.61)	0.79 (0.77-0.82)	2.52 (1.98-3.20)	0.55 (0.46-0.64)	5.31 (3.72-7.58)	57.0%
<i>Quality</i>							
High (STARD >12)	10	0.58 (0.54-0.62)	0.79 (0.75-0.82)	2.96 (2.13-4.10)	0.55 (0.44-0.69)	6.01 (3.57-10.12)	66.2%
Low (STARD ≤12)	13	0.60 (0.55-0.64)	0.79 (0.75-0.83)	2.18 (1.55-3.05)	0.53 (0.41-0.69)	4.66 (2.76-7.88)	52.4%
<i>Transplants incl/excl</i>							
Included	10	0.57 (0.52-0.62)	0.83 (0.79-0.87)	3.08 (2.00-4.73)	0.52 (0.39-0.70)	6.84 (3.29-14.23)	73.0%
Excluded	13	0.60 (0.56-0.64)	0.77 (0.73-0.80)	2.17 (1.67-2.83)	0.57 (0.46-0.69)	4.45 (3.10-6.39)	33.1%
<i>*Era</i>							
Early (pre-2002)	11	0.49 (0.45-0.54)	0.85 (0.82-0.88)	2.79 (2.11-3.68)	0.61 (0.51-0.73)	5.11 (3.26-8.02)	39.9%
Late (2002 and after)	12	0.66 (0.62-0.70)	0.73 (0.69-0.77)	2.27 (1.61-3.20)	0.48 (0.35-0.65)	5.27 (2.96-9.37)	70.1%
<i>Study size</i>							
Small (0-50)	5	0.63 (0.52-0.73)	0.78 (0.65-0.88)	2.40 (1.50-3.84)	0.53 (0.37-0.75)	6.30 (2.74-14.49)	0.0%
Medium (50-100)	10	0.66 (0.61-0.70)	0.73 (0.68-0.77)	2.14 (1.43-3.19)	0.48 (0.33-0.71)	5.01 (2.41-10.42)	74.1%
Large (>100)	8	0.53 (0.49-0.57)	0.83 (0.80-0.86)	3.04 (2.23-4.14)	0.59 (0.49-0.71)	5.42 (3.50-8.37)	53.4%
<i>Location</i>							

Location

Europe	10	0.55 (0.50-0.59)	0.84 (0.81-0.87)	3.04 (2.10-4.39)	0.52 (0.39-0.70)	6.62 (3.35-13.05)	74.6%
Rest of world	13	0.62 (0.58-0.65)	0.73 (0.69-0.77)	2.11 (1.59-2.80)	0.56 (0.46-0.68)	4.47 (3.07-6.50)	27.7%
<i>Aetiology</i>							
POD	6	0.58 (0.51-0.65)	0.89 (0.85-0.93)	4.21 (2.89-6.12)	0.48 (0.35-0.66)	10.44 (4.94-22.05)	43.1%
Non-POD	13	0.58 (0.54-0.63)	0.74 (0.69-0.78)	2.05 (1.47-2.85)	0.58 (0.45-0.76)	4.16 (2.34-7.40)	59.8%
<i>Transplant centre</i>							
Yes	15	0.56 (0.52-0.61)	0.83 (0.80-0.86)	2.89 (2.17-3.84)	0.52 (0.41-0.66)	6.54 (3.90-10.95)	61.6%
No	8	0.61 (0.57-0.65)	0.73 (0.69-0.77)	2.05 (1.42-2.94)	0.58 (0.46-0.72)	4.09 (2.52-6.62)	46.9%

B) MELD

	No. of studies	Sensitivity (95% CI)	Specificity (95% CI)	Positive LR (95% CI)	Negative LR (95% CI)	DOR (95% CI)	<i>I</i> ² (%) DOR
All	23	0.74 (0.71-0.77)	0.67 (0.64-0.69)	2.42 (1.99-2.96)	0.38 (0.31-0.46)	7.01 (5.06-9.71)	47.7%
<i>Quality</i>							
High (STARD >12)	10	0.75 (0.71-0.78)	0.62 (0.58-0.66)	2.09 (1.59-2.75)	0.41 (0.32-0.54)	5.62 (3.59-8.81)	55.9%
Low (STARD ≤12)	13	0.73 (0.69-0.77)	0.73 (0.68-0.77)	2.64 (2.22-3.13)	0.34 (0.24-0.47)	8.92 (5.59-14.23)	35.1%
<i>Transplants incl/excl</i>							
Included	10	0.70 (0.66-0.75)	0.62 (0.57-0.66)	2.11 (1.53-2.91)	0.47 (0.33-0.66)	5.49 (2.99-10.09)	62.2%
Excluded	13	0.77 (0.73-0.80)	0.70 (0.66-0.73)	2.42 (2.13-2.75)	0.34 (0.29-0.41)	7.69 (5.63-10.48)	13.7%
<i>*Era</i>							
Early (pre-2002)	11	0.74 (0.70-0.78)	0.63 (0.58-0.67)	2.19 (1.61-2.98)	0.43 (0.32-0.57)	5.68 (3.40-9.47)	54.3%
Late (2002 and after)	12	0.74 (0.71-0.78)	0.70 (0.66-0.74)	2.43 (2.12-2.79)	0.34 (0.25-0.46)	8.40 (5.68-12.42)	30.2%
<i>Study size</i>							
Small (0-50)	5	0.75 (0.65-0.83)	0.83 (0.71-0.91)	3.73 (2.06-6.75)	0.30 (0.13-0.68)	13.68 (3.94-47.51)	46.8%
Medium (50-100)	10	0.79 (0.74-0.83)	0.70 (0.65-0.75)	2.55 (2.17-2.99)	0.32 (0.24-0.42)	9.24 (6.30-13.57)	0.0%
Large (>100)	8	0.71 (0.67-0.75)	0.63 (0.59-0.67)	1.94 (1.42-2.66)	0.46 (0.36-0.60)	4.49 (2.94-6.85)	57.2%
<i>Location</i>							
Europe	10	0.73 (0.68-0.77)	0.63 (0.59-0.67)	2.17 (1.62-2.90)	0.41 (0.28-0.58)	6.13 (3.45-10.87)	65.5%
Rest of world	13	0.75 (0.72-0.79)	0.71 (0.66-0.75)	2.44 (2.11-2.84)	0.36 (0.29-0.45)	7.44 (5.46-10.14)	5.5%
<i>Aetiology</i>							

POD	6	0.80 (0.74-0.86)	0.53 (0.47-0.59)	1.74 (1.25-2.41)	0.31 (0.13-0.75)	6.55 (2.12-20.21)	70.3%
Non-POD	13	0.76 (0.72-0.80)	0.73 (0.69-0.78)	2.73 (2.29-3.26)	0.35 (0.27-0.45)	8.42 (5.98-11.88)	0.0%
<i>Transplant centre</i>							
Yes	15	0.72 (0.68-0.76)	0.64 (0.61-0.68)	2.27 (1.74-2.95)	0.41 (0.30-0.56)	6.45 (3.93-10.61)	59.0%
No	8	0.76 (0.72-0.80)	0.70 (0.65-0.75)	2.41 (2.05-2.82)	0.34 (0.30-0.41)	7.25 (5.23-10.0)	4.4%

Table 3: Subgroup analysis of the performance of A) King's College Criteria (KCC) and B) Model for End Stage Liver Disease (MELD) in predicting outcome in acute liver failure. (STARD – standards for the reporting of diagnostic accuracy studies, AALF acetaminophen related acute liver failure, NAALF non acetaminophen related acute liver failure, LR likelihood ratio, DOR diagnostic odds ratio, CI confidence interval.*midway year of study recruitment

Section and topic	Item	Description	Interpretation
Title, abstract, and keywords	1	Identify the article as a study of diagnostic accuracy (recommend MeSH heading “sensitivity and specificity”)	Title, abstract, or words contain specificity/sensitivity/PPV/NPV or words to that effect.
Introduction	2	State the research questions or aims, such as estimating diagnostic accuracy or comparing accuracy between tests or across participant groups	(as stated)
Methods: participants	3	Describe the study population: the inclusion and exclusion criteria and the settings and locations where the data were collected	(as stated)
	4	Describe participant recruitment: was this based on presenting symptoms, results from previous tests, or the fact that the participants had received the index tests or the reference standard?	Not used as not RCT and covered in 3
	5	Describe participant sampling: was this a consecutive series of participants defined by selection criteria in items 3 and 4? If not, specify how participants were further selected	(as stated - no mark if not stated)
	6	Describe data collection: was data collection planned before the index tests and reference standard were performed (prospective study) or after (retrospective study)?	One point for prospective, nil for retrospective or not stated.
Methods: test methods	7	Describe the reference standard and its rationale	Definition / calculation of KCC and MELD stated <u>or</u> reference to calculation.

	8	Describe technical specifications of material and methods involved, including how and when measurements were taken, or cite references for index tests or reference standard, or both	Mark for stating that bloods were taken (but not the biochemical assay) and clinical parameters assessed and WHEN this was done.
	9	Describe definition of and rationale for the units, cutoff points, or categories of the results of the index tests and the reference standard	Mark for papers that provide a rationale for the MELD cutoff they use (e.g. ROC calculation or reference to another paper).
Methods: statistics	12	Describe methods for calculating or comparing measures of diagnostic accuracy and the statistical methods used to quantify uncertainty (eg 95% confidence intervals)	Specifies which statistical tests and p value used.
Results: participants	14	Report when study was done, including beginning and ending dates of recruitment	(as stated)
	15	Report clinical and demographic characteristics (eg age, sex, spectrum of presenting symptoms, comorbidity, current treatments, and recruitment centre)	(as stated)
	16	Report how many participants satisfying the criteria for inclusion did or did not undergo the index tests or the reference standard, or both; describe why participants failed to receive either test (a flow diagram is strongly recommended)	(as stated)
Results: test results	17	Report time interval from index tests to reference standard, and any treatment administered between	?n/a (no time interval between index test (KCC/MELD) and reference standard (pre-defined cut-off). Timing to index test covered by (8))
	18	Report distribution of severity of disease (define criteria) in those with the target condition and other diagnoses in participants without the target condition	Range of MELD or SOFA/APACHE/equivalent scores shown for ALF population (may be entire study population).

Results: estimates	19	Report a cross tabulation of the results of the index tests (including indeterminate and missing results) by the results of the reference standard; for continuous results, report the distribution of the test results by the results of the reference standard	Table of sensitivity/specificity and/or PPV/NPV for KCC and MELD.
	20	Report any adverse events from performing the index test or the reference standard	n/a
	21	Report estimates of diagnostic accuracy and measures of statistical uncertainty (e.g. 95% confidence intervals)	Mark for showing 95% confidence intervals.
	22	Report how indeterminate results, missing responses, and outliers of index tests were handled	(as stated - n/a used if no indeterminate results, missing responses, or outliers)
	23	Report estimates of variability of diagnostic accuracy between readers, centres, or subgroups of participants, if done	Note: only multicentre studies are Taylor et al 2006, Hadem et al 2012
Discussion	24	Report estimates of test reproducibility, if done	Only included if derivation and validation cohort (Kumar et al 2012).
	25	Discuss the clinical applicability of the study findings	(as stated)

Supplementary Table S1: STARD proforma as applied to this study, n/a not available

Author	Study details			Data source		Method for converting PT to INR
	Overall aim	Study design	Outcomes assessed	For KCC	For MELD	
Pelaez-Luna et al 2006	Derive new score + compare to KCC/MELD	Retrospective cohort study	Accuracy of new scoring system (ALFIBM) vs KCC / MELD	Calculated from PPV/NPV	Calculated from PPV/NPV	Not specified (although mean +/- SD shown for PT and INR)
Taylor et al 2006	Identify prognostic markers; Derive new score + compare to KCC/MELD; Assess ALF Study Group vs UNOS cohort	Prospective observational study	Accuracy of new prognostic index vs KCC/MELD; Comparison of ALF study group vs UNOS Hep A patients	Calculated from sensitivity/specificity	Calculated from sensitivity/specificity	Not specified
Zaman et al 2006	Compare KCC/MELD	Retrospective cohort study	Accuracy of KCC/MELD	Calculated from sensitivity/specificity	Calculated from sensitivity/specificity	Not specified
Dhiman et al 2007	Compare KCC/MELD; Identify prognostic markers	Retrospective cohort study	Accuracy of prognostic indicators vs KCC/MELD	Calculated from totals/TP	Calculated from totals/TP	Not specified (PT given)
Katoonizadeh et al 2007	Compare KCC/MELD/Clichy	Retrospective cohort study	Accuracy of scores	Calculated from sensitivity/specificity	Calculated from sensitivity/specificity	Not specified
Schmidt & Larsen 2007	Compare KCC/MELD	Prospective observational study	Accuracy of KCC/MELD	Calculated from TP/spec	Calculated from TP/spec	Not specified
Wei et al 2007	Compare KCC/MELD; Description of ALF in Sweden	Retrospective cohort study	Accuracy of KCC/MELD	Calculated from sensitivity/specificity	Calculated from sensitivity/specificity	Not specified
Yantorno et al 2007	Compare KCC/MELD/Clichy	Retrospective cohort study	Accuracy of scores	Calculated from PPV/NPV	As given	Not specified
Hadem et al 2008	Compare KCC/MELD/others	Retrospective cohort study	Accuracy of scores	Calculated from sensitivity/specificity	Calculated from sensitivity/specificity	Not specified (PT given as % normal)
Yamagishi et al 2009	Compare new CT-based score to KCC/MELD	Retrospective cohort study	Accuracy of CT score vs KCC/MELD	Calculated from TP/FP	Calculated from TP/FP	Not specified
Bechmann et al 2010	Compare biomarker to KCC/MELD	Prospective observational study	Accuracy of biomarker vs KCC/MELD	As given	Calculated from sensitivity/specificity	Not specified
Kumar et al 2010	Describe ALF due to TB meds	Prospective observational study	Description of ALF due to TB meds, incl prognostic scores	Calculated from TP/FP	Calculated from sensitivity/specificity	See below
Chen et al 2010	Compare KCC/MELD/MELD Na	Retrospective cohort study	Accuracy of KCC/MELD/MELD Na	Calculated from TP/FP	Calculated from sensitivity/specificity	Not specified

Cholongitas et al 2012	Compare KCC/MELD/SOFA/APACHE II	Prospective observational study	Accuracy of KCC/MELD/SOFA/APACHE II	Calculated from TP/FP	Calculated from sensitivity/specificity	Not specified (PT given)
Craig et al 2012	Compare KCC/MELD/SOFA/MELD variants	Retrospective cohort study	Accuracy of KCC/MELD/SOFA/MELD variants	Calculated from TP/FP	Calculated from sensitivity/specificity	Not specified (PT given as seconds)
Hadem et al 2012	Describe ALF in Germany	Retrospective cohort study	Description; Accuracy of KCC/MELD/BIIE	Calculated from TP/FP	Calculated from TP/FP	Not specified
Kumar et al 2012	Derive new score + compare to KCC/MELD	Prospective observational study	Accuracy of new scoring system (ALFED) vs KCC / MELD	Calculated from TP/FP	As given / calculated from sensitivity/specificity (for validation cohort)	Not specified
Parkash et al 2012	Compare KCC/MELD	Prospective observational study	Accuracy of KCC/MELD	As given	Calculated from sensitivity/specificity	Not specified (median + IQR shown for PT and INR)
Shaikh et al 2012	Compare KCC/MELD	Retrospective cohort study	Accuracy of KCC/MELD	Calculated from TP/FP	Calculated from TP/FP	Not specified
Manka et al 2013	Compare KCC/MELD/MELD-Na variants	Prospective observational study	Accuracy of KCC/MELD/others	Calculated from sensitivity/specificity	Calculated from sensitivity/specificity	Not specified
Feng et al 2014	Compare KCC/MELD vs indocyanine green clearance test	Retrospective cohort study	Accuracy of KCC/MELD vs indocyanine green clearance test	As given	Calculated from sensitivity/specificity	Not specified
Mendizabal et al 2014	Describe ALF in Argentina, compare KCC/MELD/Clichy, + identify prognostic factors	Retrospective cohort study	Description of ALF, accuracy of KCC/MELD/Clichy, prognostic factors	Calculated from TP/FP	Calculated from TP/FP	Not specified
Bagchi et al 2015	Compare KCC/MELD vs actin-free Gc-globulin	Prospective observational study	Accuracy of KCC/MELD vs actin-free Gc-globulin	Calculated from sensitivity/specificity	Calculated from sensitivity/specificity	Not specified

Supplementary Table S2: Study design and aims with details on how 2x2 tables were calculated. For Kumar et al 2010 - Prothrombin time (PT) prolongation over control was estimated every day in each patient. It was calculated on freeze (80 C)-thawed (37 C) plasma by the manual tilt-tube technique. The clotting time in seconds was recorded after adding 200 IL thromboplastin reagent (Liquiplastin—rabbit brain thromboplastin; Tulip Diagnostic (P) Ltd., India; ISI ¼ 1.5) to 100 IL plasma. TP – true positive, FP – false positive, TN- true negative, FN – false negative, PPV – positive predictive value, NPV – negative predictive value

A) KCC

B) MELD

Figure 4 (Supplementals): Pooled diagnostic odds ratio (DOR) for for a) Kings College Criteria (KCC) and b) Model for End Stage Liver Disease (MELD)

a) KCC

b) MELD

Figure 5 (Supplementals): Deeks' Funnel Plot for a) Kings College Criteria (KCC) and b) Model for End Stage Liver Disease (MELD)