Supplementary Material 1 (Clinical Credibility):

1) Whether the risk model uses oesophagectomy specific parameters
   a. Models designed specifically for patients undergoing oesophagectomy were scored one point
2) Whether the model avoids the use of thresholds for categorisation of data
   a. Models without the use thresholds were scored one point
   b. Models using only thresholds for the assessment of age were scored as partial
   c. Models using thresholds beyond that of age were scored negatively
3) Whether the model uses data that generate outcomes prior to the time of decision making
   a. Models using only pre-operatively available data were scored one point
   b. Models relying on the use of intra-operative data (Such as intra-operative blood loss) were scored negatively
4) Whether the data input into the model is objective and reliable
   a. Models using no subjective data were scored one point
   b. Models using subjective data (Such as estimation of blood loss or patient filled health questionnaires) were scored negatively
5) Whether the model is easy to generate
   a. Models using information available on clinical history, examination or routine investigations were scored one point
   b. Models requiring investigations that may be undertaken (such as pulmonary function) were scored as partial
   c. Models needing more specialised or resource intense data were scored negatively
6) Whether the model is rendered in such a way that is understandable to the clinician
   a. Models simplistic enough for a cursory observer to understand were scored one point
   b. Models that could be understood and generated with effort were scored partially
   c. Models with complexity beyond reasonable for clinicians to understand were scored negatively
7) Whether the model stratifies risk of outcome in to clinically practically ranges
   a. Models stratifying the scores to six or more outcomes were scored one point
   b. Models stratifying outcomes to four or five categories were scored partially
   c. Models stratifying outcomes to three or fewer categories were scored negatively

References:

Minne L, Ludikhuize J, De Jonge E, De Rooij S, Abu-Hanna A. Prognostic models for predicting mortality in elderly ICU patients: a systematic review. *Intensive Care Med*. 2011 Aug;37(8):1258-68. [PMID: 21647716 DOI: 10.1007/s00134-011-2265-6]

Findlay JM, Gillies RS, Sgromo B, Marshall RE, Middleton MR, Maynard ND. Individual risk modelling for esophagectomy: a systematic review. *J Gastrointest Surg*. 2014 Aug;18(8):1532-42. [PMID: 24760219 DOI: 10.1007/s11605-014-2524-2]

Supplementary Material 2 (Methodological Quality):

1) Study participation **(8 points)**
   a. Study population description
      i. Was there a description of the study setting and period
      ii. Was there a description of the inclusion and exclusion criteria
      iii. Was there a description of the mix of surgical procedures undertaken
      iv. Was the number of patients in the study reported
      v. Did the number of patient exceed 100
      vi. Was the mortality rate reported
      vii. Were the characteristics of the patients reported including (Age/Gender/Comorbidities etc)
   b. Is the study population representative of the source population
2) Prognostic factor and outcome measurement **(4 points)**
   a. Definition of prognostic factors
      i. Was there a clear definition of all prognostic factors
      ii. Was the type of model described (Such as logistic regression model)
   b. Measurement of prognostic factors
      i. Was the number of participants with incomplete data and the handling of missing values reported
   c. Definition of the outcome
      i. Was the outcome of interest defined (Mortality/Morbidity/Complication grade)
3) Analysis **(8 points)**
   a. Appropriate analysis and description of analysis
      i. Were all evaluation measures described
      ii. Was the model building strategy described (Such as logistic regression)
      iii. Was the test method described (Train or test set/Bootstrapping/Cross-validation)
      iv. Was discrimination and validation evaluated
      v. Was a separate test set used for testing
   b. Sufficient presentation of data
      i. Was the presentation of data sufficient to assess the adequacy of analysis
      ii. Was there selective reporting of results
      iii. Was there a comparison to a standard model

Reference:

Minne L, Ludikhuize J, De Jonge E, De Rooij S, Abu-Hanna A. Prognostic models for predicting mortality in elderly ICU patients: a systematic review. *Intensive Care Med*. 2011 Aug;37(8):1258-68. [PMID: 21647716 DOI: 10.1007/s00134-011-2265-6]