Risk assessment tools validated for patients undergoing emergency laparotomy: a systematic review

C. M. Oliver¹,²,³,*, E. Walker¹,²,³, S. Giannaris³, M. P. W. Grocott²,⁴,⁵,⁶ and S. R. Moonesinghe¹,²,³

¹UCL/UCLH Surgical Outcome Research Centre (SOuRCe), 3rd Floor, Maples Link Corridor, University College Hospital, 235 Euston Road, London NW1 2BU, UK, ²National Institute of Academic Anaesthesia Health Services Research Centre, Royal College of Anaesthetists, London, UK, ³Centre for Anaesthesia, University College London, London, UK, ⁴University Hospital Southampton NHS Foundation Trust, Southampton, UK, ⁵Integrative Physiology and Critical Illness Group, Clinical and Experimental Sciences Faculty of Medicine, University of Southampton, Southampton, UK, and ⁶University Hospital Southampton NHS Foundation Trust/University of Southampton, NIHR Respiratory Biomedical Research Unit, Southampton, UK

*Corresponding author: E-mail: cmoliver@hotmail.co.uk

Abstract

Emergency laparotomies are performed commonly throughout the world, but one in six patients die within a month of surgery. Current international initiatives to reduce the considerable associated morbidity and mortality are founded upon delivering individualised perioperative care. However, while the identification of high-risk patients requires the routine assessment of individual risk, no method of doing so has been demonstrated to be practical and reliable across the commonly encountered spectrum of presentations, co-morbidities and operative procedures. A systematic review of Embase and Medline identified 20 validation studies assessing 25 risk assessment tools in patients undergoing emergency laparotomy. The most frequently studied general tools were APACHE II, ASA-PS and P-POSSUM. Comparative, quantitative analysis of tool performance was not feasible due to the heterogeneity of study design, poor reporting and infrequent within-study statistical comparison of tool performance. Reporting of calibration was notably absent in many prognostic tool validation studies. APACHE II demonstrated the most consistent discrimination of individual outcome across a variety of patient groups undergoing emergency laparotomy when used either preoperatively or postoperatively (area under the curve 0.76–0.98). While APACHE systems were designed for use in critical care, the ability of APACHE II to generate individual risk estimates from objective, exclusively preoperative data items may lead to better-informed shared decisions, triage and perioperative management of patients undergoing emergency laparotomy. Future endeavours should include the recalibration of APACHE II and P-POSSUM in contemporary cohorts, modifications to enable prediction of morbidity and assessment of the impact of adoption of these tools on clinical practice and patient outcomes.

Key words: emergency laparotomy; postoperative mortality; prognostic tool; risk adjustment; risk assessment
Introduction

Emergency laparotomy is a commonly utilized group of intra-abdominal surgical procedures performed for a variety of acute pathologies. In excess of 30,000 emergency laparotomies are performed annually in England alone, and Emergency General Surgical (EGS) admissions are considerably more numerous.1 2

Internationally reported mortality rates following emergency laparotomy range from 13 to 18% at 30 days, increasing to 25% at 24 months.3–5 This is second only to short-term mortality after emergency open repair of life-threatening ruptured abdominal aortic aneurysm (AAA).6

Reduction of the considerable morbidity and mortality after emergency laparotomy is the focus of several ongoing national and international audit and quality improvement programs, including the National Emergency Laparotomy Audit (NELA), the Australian and New Zealand Audit of Surgical Mortality, the American College of Surgeons National Surgical Quality Improvement Program (NSQIP), the Enhanced Peri-Operative Care for High-risk patients (EPOCH) study and the Dr Foster global comparators study.6–12 Central to each of these programs is the identification of high-risk patients to target perioperative interventions and augmented pathways of care.

Because patients who undergo emergency laparotomy are markedly heterogeneous, the likelihood of suffering postoperative morbidity or mortality is not evenly distributed within patient populations. The delivery of individualized care and reduction of postoperative adverse events require that both the structure and delivery of perioperative care are tailored to the needs of the individual. To this end, substantial efforts have been made to characterise high-risk patient subgroups and to identify patients at the greatest risk of death and morbidity.8 13 14

Assessment of an individual’s risk of an adverse event may be informed by clinical judgement, use of risk assessment tools, evaluation of functional capacity or plasma biomarker assay.15 Clinical judgement may vary with experience, observations of exercise tolerance are often useless in patients requiring emergency laparotomy since they are acutely unwell and evidence to support the routine use of biomarkers has yet to be established.16–18 Risk assessment tools, which incorporate clinical variables into a score or prognostic model, currently represent the most practical means of estimating risk in patients undergoing unplanned surgery, but no tool has been widely incorporated into routine practice.

Due to prevalent co-morbidities, surgical pathologies and their systemic effects and the urgency of required intervention, patients undergoing emergency laparotomy form a population distinct from those undergoing planned general surgery,19 evidenced by a higher incidence of adverse postoperative events.19–21 Therefore, while there is evidence to support the routine use of selected risk assessment tools in other clinical contexts, generalisability of the performance of these tools to patients undergoing emergency laparotomy is unknown.22–25

The objectives of this systematic review were to identify all perioperative validation studies of risk assessment tools undertaken in adult patients undergoing emergency laparotomy and to compare the reported performance and utility of the assessed tools with the aim of identifying the best tools for routine clinical use.

Prior presentation of data

Presented at the third joint meeting of the Centre for Anaesthesia, UCL’s Current Controversies in Anaesthesia and Perioperative Medicine and the Intensive Care Society of Ireland Autumn Meeting in Dingle, Ireland, September 2013.

Methods

This systematic review was registered with the PROSPERO database (CRD42014009062). Methods and reporting conform to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA), BMC and Cochrane guidelines.26–28

Definitions for the purposes of this review

Emergency
Urgent, emergent and immediately indicated surgical interventions.

Laparotomy
Open intra-abdominal surgery performed for non-aortic pathologies.

Risk assessment tool
A scoring system or prognostic model incorporating two or more variables to stratify or predict the likelihood of a specified adverse event.

Validation study
Assessment of the accuracy of one or more risk assessment tools through application to a study population. Classified as internal (application of a newly created tool to the cohort from which it was derived by practical or mathematical techniques), temporal [application of a tool to a cohort distinct in time from the derivation cohort at the institution(s) in which it was created] or external (application to patients in institutions other than that from which the tool was derived).29 30

Discrimination
How well a tool is able to discriminate between dichotomous outcomes (e.g. death and survival at 30 days) across a spectrum of risk profiles within a population of patients. Presentation as area under the receiver operator characteristic curve (AUC) provides a single, quantitative measure of the accuracy of a prognostic tool and also facilitates the comparison of dissimilar systems.31 In interpreting AUC values: >0.9, good discrimination; 0.7–0.9, moderate; and <0.7, poor.31

Calibration
How closely a prognostic model’s estimations match the observed incidence of a specified outcome across a study population. Assessed using $χ^2$ techniques, $P<0.05$ indicates that observed and expected outcomes are similar and $P<0.05$ differences are statistically significant.
Search methods and inclusion criteria

The literature search was undertaken with reference to methods of bias mitigation. Embase (a registered trademark of Elsevier B.V.) and Medline (U.S. National Library of Medicine) were searched using database-specific search terms (a complete list of search terms used in Medline is included in the Supplementary materials (S5)). Because the term emergency laparotomy is used internationally to describe a wide and varied assortment of surgical procedures and pathologies, an inclusive search strategy was adopted to achieve the aims of this systematic review.

The search was restricted to publications relating to adult humans since 1980, but was not limited to English-language publications. The last complete search was performed on March 27, 2013. The Cochrane database of systematic reviews was accessed on November 2, 2014. Secondary searching included hand-searching of references (snow-balling) and review of citation listings in Web of Knowledge (a trademark of Thomson Reuters).

Inclusion criteria
Studies assessing the discrimination of a specified outcome, presented as the AUC, by one or more risk assessment tools in adult patients undergoing emergency laparotomy were included. Studies including both emergent and elective cases were included if discrimination was reported for patients who had undergone emergency surgery.

Exclusion criteria
In order to identify useful perioperative decision-making tools, studies were excluded if no assessment of risk was made using preoperative or intraoperative data items. Validation studies confined to cohorts undergoing emergency aortic surgery and those including extra-abdominal procedures were excluded due to overt differences in patient characteristics, operative procedures and patient outcomes.

Data extraction
Data extraction was performed by the authors CO and EW and recorded directly into purpose-built tables summarising study characteristics, design quality, patient outcomes and tool performance. Differences in extracted data were discussed and consensus reached.

Extracted study characteristics included geographical region, patient cohort size and characteristics, nature of the included surgical procedures, timing of data collection (relative to emergency laparotomy) and risk assessment tools studied. Tool applicability was then classified as either general (heterogeneous and multiple subpopulation cohorts) or subpopulation specific (applicable only to cohorts defined by patient or surgical characteristics).

Extracted indicators of quality included the number of patients in the study cohort, number of institutions collecting data (single vs multicentre), timing of data collection (prospective vs retrospective), reporting of cohort baseline characteristics, reporting of inclusion criteria and excluded patients and validation methodology.

Extracted outcomes were as reported in the manuscript and included the incidence of mortality and morbidity at specified time points for identified pathologies and surgical indications.

Extracted tool performance characteristics included the AUC for a specified outcome and prognostic tool calibration and AUC 95% confidence interval (CI) where reported.

Data analysis
Decisions to pool data for meta-analysis of the performance of individual tools were informed by assessment of the homogeneity of extracted study characteristics, where overt heterogeneity of inclusion criteria, study design and patient characteristics would preclude statistical assessment of homogeneity.

Tool generalisability was determined by assessment of discrimination across dissimilar populations, including heterogeneous patient cohorts and subpopulations defined by patient or surgical characteristics.

Results
Overview
In total, 23 073 papers were identified in the primary electronic databases search, leaving 15 030 after restrictions. A further 802 papers were identified in the secondary search (Supplementary material, S1). After exclusions, 20 studies were eligible for data analysis and synthesis, assessing 25 risk assessment tools in >110 000 patients undergoing unplanned intra-abdominal surgery across 12 countries (Table 1). No similar systematic review was identified in the Cochrane Database of Systematic Reviews.

Study design and populations are summarised in Table 1. External validity was assessed in 13 studies and temporal validity in 2 studies. Internal validation techniques included split cohorts, crossover and bootstrapping.

Markers of methodological quality are presented in Table 1. Seven studies were conducted across multiple institutions, validation cohort size varied from 49 to >68 000 patients, and data collection was performed entirely prospectively in 15 studies and demographic data were presented in 19 papers. Reporting of inclusion criteria, exclusions and surgical procedures was universally adequate. Reporting of calibration and AUC 95% CIs was inconsistent. Statistical comparison of tool discrimination was reported in only one study.

Short-term (30-day or inpatient) mortality endpoints were reported in all studies; 30-day mortality was reported in 11 studies and inpatient mortality in 9 studies. The 30-day mortality ranged from 9 to 27% and inpatient mortality from 3 to 26%, varying by operative procedure, surgical indication and patient age (Table 2). Other identified endpoints included postoperative morbidity and complications. These were infrequently reported (Table 1).

Overt heterogeneity of study design, patient characteristics and presented outcomes precluded meta-analysis, necessitating a qualitative approach to tool comparison (Table 1).

Risk assessment tools
Thirteen general tools and 12 subpopulation-specific tools were assessed in the identified studies (Table 1). In addition to the Physiological and Operative Severity Score for the enumeration of Mortality and morbidity (POSSUM), several previously derived POSSUM systems and novel coefficients were assessed.

Many tools were not created for the purpose of perioperative risk assessment, including tools for critical care [Acute Physiology and Chronic Health Evaluation II (APACHE II), APACHE III, the Simplified Acute Physiology Score II (SAPS II) and the Mortality Probability Model II (MPM II)], the Early Warning Score and ‘sepsis score and the POSSUM systems, which were created for comparative audit.

Two tools specific to emergency laparotomy were identified: the preoperative and perioperative NSQIP Emergency Laparotomy models. These were assessed in a single, large internal validation study.

Tool characteristics are summarised in Supplementary material S2. There was notable variation in the required number of data items (ranging from the one composite measure of the...
<table>
<thead>
<tr>
<th>First Author</th>
<th>Year</th>
<th>Region</th>
<th>Cohort size</th>
<th>Single/multi centre</th>
<th>Data acquisition</th>
<th>Baseline data reported</th>
<th>Inclusion criteria</th>
<th>Exclusions</th>
<th>Models assessed</th>
<th>Outcome</th>
<th>Validation methodology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Al-Temimi</td>
<td>2012</td>
<td>USA</td>
<td>37 553</td>
<td>Multi</td>
<td>Prospective</td>
<td>Yes</td>
<td>&gt;16 years, EL for general surgical indications or mesenteric insufficiency</td>
<td>Missing data, urgent or vascular surgery, laparoscopic procedure converted to open</td>
<td>1) EL perioperative model, 2) EL preoperative model</td>
<td>30-day mortality</td>
<td>Internal: crossover by year</td>
</tr>
<tr>
<td>Biondo</td>
<td>2000</td>
<td>Spain</td>
<td>55</td>
<td>Single</td>
<td>Prospective</td>
<td>Yes</td>
<td>Consecutive patients, emergency surgery for distal colonic peritonitis</td>
<td>'Complicated non-specific inflammatory disease of the colon'</td>
<td>1) Colonic peritonitis severity score</td>
<td>Inpatient mortality</td>
<td>Internal: split</td>
</tr>
<tr>
<td>Biondo</td>
<td>2006</td>
<td>Spain</td>
<td>156</td>
<td>Single</td>
<td>Prospective</td>
<td>Yes</td>
<td>Clinical diagnosis of peritonitis, emergency surgery</td>
<td>None declared</td>
<td>1) Colonic Peritonitis Severity Score, 2) Mannheim Peritonitis Index</td>
<td>Inpatient mortality</td>
<td>Temporal</td>
</tr>
<tr>
<td>Buck</td>
<td>2012</td>
<td>Denmark</td>
<td>117</td>
<td>Multi</td>
<td>Prospective</td>
<td>Yes</td>
<td>Surgical treatment for perforated peptic ulcer</td>
<td>Pregnant or breastfeeding, malignant ulcers, perforation of another organ</td>
<td>1) Boey score, 2) ASA-PS, 3) APACHE II, 4) sepsis score</td>
<td>30-day mortality, septic shock</td>
<td>External</td>
</tr>
<tr>
<td>Ertan</td>
<td>2008</td>
<td>Turkey</td>
<td>102</td>
<td>Single</td>
<td>Retrospective</td>
<td>Yes</td>
<td>Emergency colorectal surgery for complications of colorectal carcinoma</td>
<td>Uncertain diagnosis, insufficient data</td>
<td>1) APACHE III, 2) MPM II, 3) CR-POSSUM</td>
<td>30-day mortality</td>
<td>External</td>
</tr>
<tr>
<td>Ferjani</td>
<td>2007</td>
<td>UK</td>
<td>158</td>
<td>Single</td>
<td>Prospective</td>
<td>Yes</td>
<td>Consecutive patients, histologically confirmed colorectal cancer, abdominal surgery to remove primary tumour</td>
<td>Laparoscopic surgery</td>
<td>1) ACPGBI, 2) POSSUM, 3) P-POSSUM, 4) CR-POSSUM</td>
<td>30-day mortality</td>
<td>External</td>
</tr>
<tr>
<td>Garcea</td>
<td>2010</td>
<td>UK</td>
<td>280</td>
<td>Single</td>
<td>Retrospective</td>
<td>Yes</td>
<td>EL for suspected perforation of a viscus/obstruction/fulminant colitis/upper gastrointestinal bleeding/surgery for strangulated ventral or groin hernia</td>
<td>None declared</td>
<td>1) Early Warning Score, 2) ASA-PS, 3) POSSUM, 4) APACHE II</td>
<td>ICU and total LOS, inpatient mortality,</td>
<td>External</td>
</tr>
<tr>
<td>Goffi</td>
<td>1999</td>
<td>Italy</td>
<td>49</td>
<td>Single</td>
<td>Prospective</td>
<td>Yes</td>
<td>Major emergency operations, including trauma</td>
<td>None declared</td>
<td>1) APACHE II</td>
<td>30-day mortality</td>
<td>External</td>
</tr>
<tr>
<td>Last name</td>
<td>Journal</td>
<td>Country</td>
<td>N</td>
<td>Study Type</td>
<td>Design</td>
<td>Age Criteria</td>
<td>Outcome Measures</td>
<td>Risk Assessment Tools</td>
<td>Internal Validity</td>
<td>External Validity</td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>---------</td>
<td>---------</td>
<td>---</td>
<td>------------</td>
<td>--------</td>
<td>--------------</td>
<td>-------------------</td>
<td>---------------------</td>
<td>-------------------</td>
<td>------------------</td>
<td></td>
</tr>
<tr>
<td>Kermani</td>
<td>2013 USA</td>
<td>68 344</td>
<td>Multi Retrospective Yes</td>
<td>ICD-9 coding: &gt;18 years, admitted non-electively, colectomy</td>
<td>Missing data</td>
<td>1) Practical mortality risk score for emergent colectomy</td>
<td>Inpatient mortality</td>
<td>Internal: crossover (10-fold k-partitions)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Koc</td>
<td>2007 Turkey</td>
<td>75 Single Prospective Yes</td>
<td>Emergency surgery for perforated peptic ulcer</td>
<td>Surgery for perforated ulcer at the site of a previous anastomosis, malignancy</td>
<td>Missing data</td>
<td>1) APACHE II, 2) APACHE III, 3) SAPS II, 4) MPM II</td>
<td>30-day mortality</td>
<td>External</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kologlu</td>
<td>2001 Turkey</td>
<td>473 Single Retrospective No</td>
<td>Operation for intra-abdominal infection without continuous postoperative peritoneal lavage</td>
<td>Missing data, percutaneous drainage of intra-abdominal abscess, uncomplicated appendicitis, uncomplicated cholecystitis, planned repeat laparotomy</td>
<td>1) Mannheim Peritonitis Index, 2) Peritonitis Index of Altona, 3) Combined peritonitis score</td>
<td>Inpatient mortality</td>
<td>External</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kulkarni</td>
<td>2007 India</td>
<td>50 Single Prospective Yes</td>
<td>Peritonitis due to perforation of hollow viscus</td>
<td>Blunt abdominal trauma with associated solid organ/vascular/neurological injury or fracture</td>
<td>1) APACHE II</td>
<td>Inpatient mortality</td>
<td>External</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kwok</td>
<td>2011 USA</td>
<td>372 Multi Prospective Yes</td>
<td>&gt;80 years, CPT code: emergency colectomy</td>
<td>Laparotomy resulting in ‘diversion only’, i.e. without colonic resection, missing data</td>
<td>1) Targeted risk prediction score, 2) ASA-PS, 3) Surgical Risk Scale, 4) ACS colorectal surgery risk calculator</td>
<td>30-day mortality</td>
<td>Temporal</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lohsiriwat</td>
<td>2008 Thailand</td>
<td>152 Single Prospective Yes</td>
<td>Emergency surgery for perforated peptic ulcer</td>
<td>Perforated gastric cancer</td>
<td>1) Boey score, 2) ASA-PS, 3) Mannheim Peritonitis Index</td>
<td>30-day mortality, complications</td>
<td>External</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Merad</td>
<td>2012 France</td>
<td>575 Multi Prospective Yes</td>
<td>&gt;16 years, major digestive surgery</td>
<td>&gt;1 missing P-POSSUM value</td>
<td>1) P-POSSUM</td>
<td>Inpatient mortality</td>
<td>External</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moller</td>
<td>2012 Denmark</td>
<td>2668 Multi Prospective Yes</td>
<td>Surgical treatment of benign gastric or duodenal perforated peptic ulcer</td>
<td>Malignant peptic ulcer</td>
<td>1) Peptic ulcer prediction score, 2) ASA-PS, 3) Boey score</td>
<td>30-day mortality</td>
<td>Internal: bootstrapping for PULP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
ASA physical status classification (ASA-PS) to the 41 variables comprising the NSQIP perioperative Emergency Laparotomy model, the preoperative availability of data items, complexity of calculation and the requirement for subjective interpretation of clinical data items.

General tools
Of the 13 identified general tools (Table 3), the most frequently assessed were APACHE II (7 studies), ASA-PS (5 studies) and P-POSSUM (4 studies). Each tool was assessed in both heterogeneous patient cohorts and subpopulations defined by demographic characteristics, surgical indication or operative procedure.

The ability of APACHE II to discriminate between short-term death and survival was moderate to good (AUC 0.76–0.98) when applied to heterogeneous cohorts undergoing unplanned intra-abdominal surgery for a variety of indications including peritonitis, colorectal malignancy and perforated peptic ulcer. Notably, APACHE II was scored using exclusively preoperative data in four studies: at hospital admission in three studies and on booking for a surgical theatre in one study.

Discrimination between short-term outcomes by the ASA-PS was moderate or good in patient cohorts undergoing emergency laparotomy or repair of perforated peptic ulcer (AUC 0.81 and AUC 0.73–0.91, respectively), but poor in an elderly cohort (AUC 0.66).

Notably, APACHE II was scored using exclusively preoperative data in four studies: at hospital admission in three studies and on booking for a surgical theatre in one study.

Discrimination of 30-day survival by NSQIP emergency laparotomy models was moderately good (AUC 0.87–0.88) in a single internal validation study.

Reporting of prognostic tool calibration was poor and, where reported, performance was inconsistent (Table 3).

Subpopulation-specific tools
A variety of indication-specific and co-morbidity-specific tools were identified (Supplementary materials (S3)).

Colorectal surgery. Of six tools specific to colorectal surgery, CR-POSSUM alone was assessed in multiple studies. Discrimination of 30-day outcome after unplanned surgery for colorectal cancer by CR-POSSUM was moderate or poor (AUC 0.65, 0.72).

Peritonitis. Of four identified studies of tools specific to peritonitis, the Mannheim Peritonitis Index (MPI) and Peritonitis Index of Altona II (PIA II) underwent multiple assessments of external validity. Discrimination of short-term outcome by the MPI and PIA II...
Table 3 Discrimination and calibration of general tools. GOF, goodness of fit; PPU, perforated peptic ulcer; H-L, Hosmer-Lemeshow statistic

<table>
<thead>
<tr>
<th>Tool</th>
<th>Total patients</th>
<th>Region</th>
<th>First author (year)</th>
<th>Cohort size</th>
<th>Surgery subtype</th>
<th>Inclusion criteria</th>
<th>Primary endpoint</th>
<th>Incidence of primary endpoint</th>
<th>AUC (95% CI)</th>
<th>Calibration P-value</th>
<th>H-L value (unless stated)</th>
</tr>
</thead>
<tbody>
<tr>
<td>APACHE II</td>
<td>944</td>
<td>Turkey</td>
<td>Ertan (2008)</td>
<td>102</td>
<td>General Surgery</td>
<td>Colorectal cancer</td>
<td>30-day mortality</td>
<td>17%</td>
<td>0.78</td>
<td>0.49</td>
<td>4.448</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Italy</td>
<td>Goffi (1999)</td>
<td>49</td>
<td>General Surgery</td>
<td>Mixed</td>
<td>30-day mortality</td>
<td>20%</td>
<td>0.87</td>
<td>0.63</td>
<td>χ² quintiles</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Europe</td>
<td>Ohmann (1993)</td>
<td>271</td>
<td>General Surgery</td>
<td>Peritonitis</td>
<td>30-day mortality</td>
<td>21%</td>
<td>0.87</td>
<td>&gt;0.05</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Turkey</td>
<td>Koc (2007)</td>
<td>75</td>
<td>PPU Repair</td>
<td>PPU</td>
<td>30-day mortality</td>
<td>11%</td>
<td>0.87</td>
<td>0.007</td>
<td>17.58</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Denmark</td>
<td>Buck (2012)</td>
<td>117</td>
<td>PPU Repair</td>
<td>PPU</td>
<td>30-day mortality</td>
<td>17%</td>
<td>0.76</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>UK</td>
<td>Garcea (2010)</td>
<td>280</td>
<td>General Surgery</td>
<td>Mixed</td>
<td>Inpatient mortality</td>
<td>15%</td>
<td>0.76</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>India</td>
<td>Kulkarni (2007)</td>
<td>50</td>
<td>General Surgery</td>
<td>Perforative peritonitis</td>
<td>Inpatient mortality</td>
<td>16%</td>
<td>0.98</td>
<td></td>
<td></td>
</tr>
<tr>
<td>APACHE III</td>
<td>177</td>
<td>Turkey</td>
<td>Ertan (2008)</td>
<td>102</td>
<td>General Surgery</td>
<td>Colorectal cancer</td>
<td>30-day mortality</td>
<td>17%</td>
<td>0.77</td>
<td>0.9</td>
<td>2.208</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Turkey</td>
<td>Koc (2007)</td>
<td>75</td>
<td>PPU Repair</td>
<td>PPU</td>
<td>30-day mortality</td>
<td>11%</td>
<td>0.84</td>
<td>0.01</td>
<td>15.08</td>
</tr>
<tr>
<td>ASA-APS</td>
<td>3589</td>
<td>USA</td>
<td>Kwok (2011)</td>
<td>372</td>
<td>General Surgery</td>
<td>&gt;80 years: colectomy</td>
<td>30-day mortality</td>
<td>26%</td>
<td>0.66</td>
<td>0.14</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Denmark</td>
<td>Buck (2012)</td>
<td>117</td>
<td>PPU Repair</td>
<td>PPU</td>
<td>30-day mortality</td>
<td>17%</td>
<td>0.73</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Denmark</td>
<td>Moller (2012)</td>
<td>2668</td>
<td>PPU Repair</td>
<td>PPU</td>
<td>30-day mortality</td>
<td>27%</td>
<td>0.78</td>
<td>(0.76–0.80)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>UK</td>
<td>Garcea (2010)</td>
<td>280</td>
<td>General Surgery</td>
<td>Mixed</td>
<td>Inpatient mortality</td>
<td>15%</td>
<td>0.81</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Early Warning Score</td>
<td>280</td>
<td>Thailand</td>
<td>Lohsiriwat (2008)</td>
<td>152</td>
<td>PPU Repair</td>
<td>PPU</td>
<td>30-day mortality</td>
<td>9%</td>
<td>0.91</td>
<td>(0.85–0.95)</td>
<td></td>
</tr>
<tr>
<td>Emergency Laparotomy</td>
<td>37 553</td>
<td>USA</td>
<td>Al-Temimi (2012)</td>
<td>37 553</td>
<td>General Surgery</td>
<td>Mixed</td>
<td>30-day mortality</td>
<td>14%</td>
<td>0.88</td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td>perioperative model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emergency Laparotomy</td>
<td>37 553</td>
<td>USA</td>
<td>Al-Temimi (2012)</td>
<td>37 553</td>
<td>General Surgery</td>
<td>Mixed</td>
<td>30-day mortality</td>
<td>14%</td>
<td>0.87</td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td>preoperative model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mortality Probability</td>
<td>177</td>
<td>Turkey</td>
<td>Ertan (2008)</td>
<td>102</td>
<td>General Surgery</td>
<td>Colorectal cancer</td>
<td>30-day mortality</td>
<td>17%</td>
<td>0.71</td>
<td>0.46</td>
<td>7.736</td>
</tr>
<tr>
<td>Model (MPM) II</td>
<td></td>
<td>Turkey</td>
<td>Koc (2007)</td>
<td>75</td>
<td>PPU Repair</td>
<td>PPU</td>
<td>30-day mortality</td>
<td>11%</td>
<td>0.98</td>
<td>0.99</td>
<td>1.36</td>
</tr>
<tr>
<td>POSSUM</td>
<td>438</td>
<td>UK</td>
<td>Ferjani (2008)</td>
<td>158</td>
<td>General Surgery</td>
<td>Colorectal cancer</td>
<td>30-day mortality</td>
<td>20%</td>
<td>0.63</td>
<td>(0.55–0.70)</td>
<td>0.037</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UK</td>
<td>Garcea (2010)</td>
<td>280</td>
<td>General Surgery</td>
<td>Mixed</td>
<td>Inpatient mortality</td>
<td>15%</td>
<td>0.81</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Continued...
was inconsistent for patients undergoing general surgery for peritonitis (AUC 0.73, 0.97 and 0.69, 0.95, respectively). 36 44 53 54

**Perforated peptic ulcer (PPU).** Of two tools specific to PPU, the Boey score alone was assessed in multiple studies. Discrimination of outcome at 30 days by the Boey score was moderate or poor (AUC 0.63–0.86), but in discrimination of development of postoperative septic shock and complications, performance was moderate (AUC 0.72 and 0.80, respectively) (Supplementary material S4).37 38 49

**Discussion**

Risk assessment tools can support clinical judgement in determining appropriate treatment plans, informing consent and tailoring perioperative care and may also be used in risk adjustment, thus supporting quality improvement.

The objectives of this systematic review were to identify all validation studies of risk assessment tools in cohorts of adults undergoing emergency laparotomy and to compare the performance and utility of these tools in order to determine which are best suited to perioperative clinical practice.

Despite the publication of guidance, study design and quality was variable, with some studies not meeting basic criteria.55 Calibration was infrequently reported.

Short-term mortality after emergency laparotomy varied considerably and was greatest in patient cohorts defined by increased age (Table 2).

The most frequently assessed tools were not created for the perioperative prediction of individual risk. APACHE II, which was created for assessment of the severity of critical illness, demonstrated the most consistent accuracy across a variety of patient cohorts undergoing emergency laparotomies when used either preoperatively or postoperatively. In contrast with many other tools, APACHE II also has the advantage that it may be used to generate individual percentage risk estimates using objective clinical data items routinely available in this setting (Supplementary material S2).

**Performance**

Because patients who require emergency laparotomy are markedly heterogeneous, the capacity to tolerate a cascade of acute surgical pathology, massive surgical insult and resulting postoperative organ dysfunction varies between individuals. Furthermore, underlying surgical pathologies may not be apparent prior to surgery. For a tool to be useful for emergency laparotomy it must therefore be both applicable to and accurate across the spectrum of patient characteristics, surgical pathologies and operative procedures encountered in clinical practice. One general tool is therefore preferable to multiple subpopulation-specific tools.

When determining the best tool for emergency laparotomy, demonstration of satisfactory and consistent performance is essential and must precede these other considerations.29–31

**Comparisons of tool performance**

Assessment of external validity provides the best measure of a tool’s generalisability and the performance of tools that are validated only internally may not be replicable in external patient populations due to factors including overfitting.50 55 Of the identified general risk assessment tools, only seven were assessed in multiple patient cohorts (Table 3). The two identified tools that
were developed specifically for emergency laparotomy were assessed in a single large database cohort and validated only internally. For those tools that were assessed in multiple cohorts, pooling of data for head-to-head comparisons of discriminatory performance was not feasible due to significant heterogeneity of both study design and patient cohorts and the infrequent reporting of AUC CIs and within-study statistical comparisons of performance.

Demonstration of prognostic tool calibration is essential both for clinicians to have confidence in estimates of risk and for the process of risk adjustment. It is therefore notable that calibration was infrequently reported in the identified validation studies of prognostic tools.

Disparities were evident in the discriminatory performance (APACHE II, 0.76–0.98; ASA-PS, 0.66–0.91; P-POSSUM, 0.65–0.82) and calibration of the most frequently assessed tools. This variation is most likely to reflect differences in study cohorts and methodologies, poor generalisability of performance to some patient subgroups and poor reliability of subjectively scored tools. However, no identified tool incorporates measures of organisational structure and processes of care or of geographical variations in patient-level risk, which may account for some of the variation observed.56

The reduced accuracy of the ASA-PS and P-POSSUM in patients with colorectal malignancy and older people suggests that the performance of these tools is not generalisable to these subgroups of patients.

The reliability of the output of risk assessment tools may be reduced by subjective interpretation of data items and inconsistent application, which may account for some of the variation observed in the performance of the ASA-PS and P-POSSUM. Both tools require the interpretation and scoring of clinical data items (including chest X-ray and ECG),42 the POSSUM system ‘multiple procedures’ item is variably interpreted by clinicians and linear and exponential analyses are variably used in the calculation of percentage predicted risk with POSSUM systems.57

APACHE II was created for the assessment of risk in critical care admissions,58 incorporating physiological parameters, markers of chronic disease and age. The comparatively good performance of APACHE II in these patients undergoing emergency laparotomy may therefore reflect the associations of age, magnitude of systemic insult and relevant co-morbidities with adverse postoperative outcomes in such cohorts of patients.

**Tool performance in core subgroups**

Complications of intra-abdominal malignancy are a common indication for emergency laparotomy,13 the incidence of colorectal cancer in older people is increasing32 and outcomes after emergency laparotomy vary with age and the timing of presentation with malignancy.61 Because elderly patients and individuals with colorectal malignancy thus represent core subgroups of patients undergoing emergency laparotomy, it is essential that the performance of tools for emergency laparotomy is generalisable to these patients.

While the significance of the observation is uncertain on the basis of the data analysed, it is therefore notable that discrimination of outcome by APACHE II in a cohort of patients with colorectal cancer was not evidently reduced (AUC 0.78), whereas P-POSSUM and CR-POSSUM discriminated less well in patients with colorectal cancer; similarly the accuracy of the ASA-PS was reduced in a cohort of elderly patients.39 50–52

Adding weight to reports that existing scoring systems are inaccurate and unreliable in elderly cohorts undergoing emergency surgery, it is notable that in a cohort of elderly patients undergoing emergency colectomy, neither age-specific nor general tools demonstrated adequate discrimination.50 62 63

**Tool characteristics**

Even if consistent and generalisable performance can be demonstrated, a tool is unlikely to be adopted into routine clinical practice if it is unwieldy. This may be true of tools requiring numerous data points, exacting a high burden of data collection or due to the complexity of required calculations.

**Utility**

The ASA-PS requires only one composite variable, APACHE II requires 12 data points that are routinely available for many patients undergoing emergency laparotomy and P-POSSUM requires 18 data points, whereas the NSQIP emergency laparotomy models require a minimum of 39 variables (Supplementary material S2). Tools that require numerous data items might be expected to better capture all relevant risk factors, but performance data did not suggest better discrimination by complex tools (Table 3).

Percentage estimates of individual risk may be used to inform a variety of shared and clinical decisions, including consent for emergency laparotomy, clinician seniority and the location of postoperative care.3–12 The ASA-PS lacks this capability, but such estimates may be generated with prognostic tools including P-POSSUM and APACHE II. And while logarithmic equations are required, the widespread availability of online and ‘app’-based calculators means that bedside estimation of individual risk is feasible and could be incorporated into routine clinical practice.

**Timing of risk assessment**

The accuracy of tools that require intra-operative and post-operative data items may not be superior to tools that require only preoperative data.49 Indeed, accuracy increased only negligibly with the inclusion of intra-operative data (AUC 0.87 and 0.88, respectively) in the NSQIP emergency laparotomy tools derivation study.4

In clinical practice, the preoperative availability of individualised estimates of risk will better inform shared decisions than estimates that are available only postoperatively. It is therefore notable that while the accuracy of APACHE II estimates calculated using only preoperative data was comparable with studies using perioperatively or postoperatively collected data (AUC 0.76–0.98 and 0.78–0.87 respectively), P-POSSUM-predicted risk was not calculated preoperatively in any of the studies identified.35 37 44–48 Preoperative performance of P-POSSUM may therefore have been overstated by the findings of this review.

**Limitations**

Few of the identified tools were assessed in multiple validation studies and many cohorts were small and drawn from single centres, limiting both statistical power and assessment of the generalisability of tool performance.

Only studies reporting tool discrimination as AUC were included in this systematic review. While the benefits of doing so are widely accepted, a small proportion of excluded studies used alternative methods of assessing and reporting tool performance in patient cohorts undergoing emergency laparotomy. Quantitative comparative analysis of tool performance (discrimination
and calibration) was not feasible due to the significant heterogeneity of study design and cohort characteristics, poor reporting and infrequent within-study statistical comparisons of tool performance.\textsuperscript{33}

Organisational factors in the delivery of health care have undergone considerable change over recent years. Because this review includes studies from the 1990s, the reported accuracy and calibration of identified tools may not describe performance in contemporary patient cohorts.\textsuperscript{55}

Bias identification and mitigation was limited by variable reporting of study methodology, potential for preferential publication of studies with positive findings and the predominance of English-language papers in the accessed electronic databases.\textsuperscript{56}

Conclusions and direction of future research
The identification of high-risk patients for targeted perioperative interventions and augmented pathways of care is central to current initiatives to improve patient outcomes after emergency laparotomy. Parallel initiatives include ensuring that the care of all patients meets agreed standards, that novel perioperative interventions continue to be developed\textsuperscript{57} and that, where possible, individuals at risk of emergency presentation are identified, counselled and managed before the acute episode.

Comparison of the performance of risk assessment tools for emergency laparotomy was not possible with existing data, highlighting the need for consensus in the reporting of perioperative outcomes.\textsuperscript{68} However, due to consistent performance across patient subgroups and its capacity to generate individual preoperative risk estimates using routinely available objective data items, APACHE II and P-POSSUM in large, well-conducted contemporary studies in heterogeneous emergency general surgical cohorts, assess the implications of dynamic risk scores (e.g. preoperative and postoperative scores), develop the ability to accurately predict postoperative morbidity due to associations with prolonged excess mortality and assess the impact of adoption of these tools on patient outcomes and clinical practice.\textsuperscript{79 50 69}

Finally, it should be recognised that in addition to the numerous potential benefits to patients of individual risk prediction, the implementation of nationally coordinated methods for the assessment of individual risk, as is true of the National Emergency Laparotomy Audit (NELA), the Intensive Care National Audit and Research Centre (ICNARC) and the American College of Surgeons (ACS) NSQIP augments and promotes the collection of risk data, health services research and quality improvement initiatives.

Author’s contributions
Study design, search strategy development, first draft and revisions: C.M.O., S.R.M.
Data extraction and analysis: C.M.O., E.W., S.G.
Critical revision of the manuscript: C.M.O., E.W., S.G., M.G., S.R.M.

Supplementary material
Supplementary material is available at British Journal of Anaesthesia online.

References
12. The Dr Foster Global Comparators Study. Available from http://drfosterintelligence.co.uk/global-comparators/

Handling editor: J. G. Hardman