The performance of organ dysfunction scores for the early prediction and management of severity in acute pancreatitis: an exploratory phase diagnostic study

Short title: Organ dysfunction scores in acute pancreatitis

James M Mason DPhil[1]

Srinivasan Balachandra FRCS,

Benoy I Babu MRCS,

Anil Bagul MRCS

Ajith K Siriwardena MD FRCS.

Regional Hepatobiliary Surgery Unit,
Manchester Royal Infirmary
Oxford Road
Manchester M13 9WL, UK.

and

[1]School of Medicine and Health
Durham University
University Boulevard
Stockton-on-Tees TS17 6BH, UK.

This study was presented in the Moynihan Prize session of the Annual Meeting of the Association of Surgeons of Great Britain and Ireland, Manchester 2007 and published in abstract form in the British Journal of Surgery [*Br J Surg* 2007:**94**(S2);2].

Correspondence to:

Professor Ajith K Siriwardena MD FRCS
Hepatobiliary Surgery Unit
Manchester Royal Infirmary
Tel: 0161-276-4250
Fax:0161-276-4530
e-mail: ajith.siriwardena@cmmc.nhs.uk

Conflicts of interest: none.

Financial support: none.

Abstract

Objective

Severity stratification in acute pancreatitis is needed both to guide clinical practice and recruit patients into studies.

Methods

Patients with a clinical acute pancreatitis (N=181) were recruited prospectively, and organ dysfunction scores (LODS, MODS and SOFA), C-reactive protein (CRP) and APACHE II scores were collected. Patients who died or used critical care (level II/III) during admission were classed as 'severe' cases. The ability of tests to accurately select patients was assessed.

Results

The ability of test measures to accurately select severe cases within the first 24 hours of admission was inadequate. At 24 hours, for a LODS score ≥ 1, sensitivity of 90% and specificity of 68% equated to a positive predictive value of 35%: only one in three patients selected for enhanced care would subsequently prove to be a severe case, although only 10% of severe cases would be missed. The use of multiple tests is unlikely to be helpful given the degree of correlation between measures.

Conclusion

Current tissue damage and organ dysfunction scoring methods are inadequate to select patients to guide subsequent care or for involvement in studies. New, better methods are necessary.

[Abstract word count: 189].

Key words: Acute pancreatitis, organ dysfunction, severity stratification.

Introduction

The 1992 Atlanta consensus conference defined two broad categories of acute pancreatitis – mild and severe[1]. Although there is a broad spectrum of disease severity in this illness[2], the distinction was relevant in terms of clinical management and remains pertinent today as newer interventions should be targeted at individuals with severe acute pancreatitis. However, two limitations to do with definition and timing are apparent in the original Atlanta classification.

Severe acute pancreatitis was defined as pancreatitis with organ failure and/or local complications[1]. Organ dysfunction is 'transient' in some patients[3] – and these patients with transient organ dysfunction often have a mild clinical course with early recovery and discharge from hospital whereas strict application of Atlanta criteria would demand that they be noted as severe. Although the majority of patients with obvious mild or severe disease will be clearly defined, allocation is less clear for individuals with transient severe disease, giving rise to scope for inter-observer variation: such variation reduces comparability between units in activity levels and health outcomes, when analysed by severity.

A major limitation of the Atlanta classification is that it provides post-episode classification and was never intended to provide prognostic information at the point of (or soon after) admission. There is a major need to differentiate high and low risk patients at, or soon after, admission both to optimise subsequent clinical care and to select consistently high risk patients for new studies. Failure to accurately predict severe disease may lead to confounding within and between studies due to a varying proportion of patients with mild illness[4].

The goal of early patient classification is to identify all patients accurately into severe and mild categories. Thus the utility of any method of classification is found in its test sensitivity and sensitivity (see Table 1) – the proportion of disease accurately classified. In addition, to

efficiently select patients for early critical management, a test must deliver a high positive predictive value or valuable critical care services are occupied unnecessarily.

Patient identification for selection into studies has a different emphasis, since the objective is to identify a group of patients who are consistently at high risk or low risk. It may be possible for a test to find such group although it may not make efficient use of all potential participants. Thus the utility of selection for studies is determined particularly by the test positive or negative predictive values, since these determine the homogeneity of the group selected.

Early identification of severe acute pancreatitis patients remains difficult despite a large body of research exploring biological prognostic markers[5-7], multiple factor scoring systems[8,9] or permutations of these[7]. As severe acute pancreatitis is accompanied by a systemic inflammatory response and multiple organ dysfunction[10], useful prognostic information may be gained from organ failure scoring systems currently used in critical care. Systems such as the logistic organ dysfunction score (LODS)[11] and the Marshall organ dysfunction score (MODS)[12] are well validated for prognostic use in critical care populations, although their performance specifically in acute pancreatitis is less well understood.

Seeking to improve the early management of acute pancreatitis, this paper explores the performance of scoring systems derived and validated in critical care medicine to categorize severity in acute pancreatitis. Organ dysfunction scoring in this context has several theoretical attractions: the score is completed by assessing a relevant and comprehensive biologic dataset; and, these data inform the need for critical care admission in addition to identifying patients at risk of adverse outcome.

Methods

Patients and setting

The study was undertaken in a Manchester Royal Infirmary, England - a university teaching hospital serving an urban population of mixed ethnicity. A consecutive series of 181 patients, with a clinical diagnosis of acute pancreatitis (AP) presenting from February 2001 to November 2004, was enrolled prospectively into the study. The clinical diagnosis of AP was based on a combination of acute abdominal pain, three-fold elevation of serum amylase and appropriate clinical features. Patients with known chronic pancreatitis and those tertiary transfers of patients with on-going pancreatic necrosis were excluded from the study.

Data collection

Patient-level data included demographics, ætiology of acute pancreatitis, duration of inpatient stay (by level of dependency), surgical/radiological interventions and mortality. APACHE II, Logistic Organ Dysfunction Score (LODS), Marshall Organ dysfunction score (MODS) and Sequential Organ Failure Score (SOFA) were estimated at admission, at 24 and 48 hours and 7 days. Day 7 data were only collected in those individuals who were still in-patients. Additionally C-reactive protein (CRP) was measured at 48 hours. The hepatic score component of LODS was omitted to avoid potential bias in patients with gallstone-related disease. Ætiologies were categorised for the purposes of analysis into: gallstone, alcohol-related, idiopathic, post-endoscopic retrograde cholangiopancreatography (ERCP) and other.

Data analysis

In-hospital death or use of critical care stay (level II and level III support)[14] was selected as the principal endpoint against which organ dysfunction scoring systems were tested. This endpoint is the most objective marker of severe disease available. Duration of hospital stay was also explored initially but was considered prone to variation between healthcare systems for reasons not directly related to disease severity[15].

Receiver operator curves were generated for each measure with area under curve (AUC) estimates being interpreted as a comparative measure of the potential usefulness of each test, and test accuracy being compared at visually optimal thresholds. Extrinsic measures (sensitivity and specificity) were used to assess the efficiency of the classification of cases; intrinsic measures (predictive values) were used to assess the adequacy of selection of cases by severity (see Table 1).

Ethical approval

Data collection for this study was under the auspices of an on-going evaluation of patients with acute pancreatitis and with institutional review board approval.

Statistical analyses

This analysis is exploratory (hypothesis generating) and concerned with estimating the accuracy of potential candidate tests to identify high and low risk patients. Consequently AUC estimates are reported with 95% confidence intervals, and no adjustment is made for multiple comparisons. Data were analysed using SPSS release 15.

Results

Ætiology and episode severity

The underlying ætiology of acute pancreatitis in this cohort was: gallstones 96 (53%), alcohol-related 42 (23%), idiopathic 27 (15%), post-ERCP 10 (6%) and 6 (3%) others. The mean age was 52 (SD20, range 17 to 87), and 48% were male. In total, 29 of 181 patients (16%) were classified by as having a severe episode as defined by death or need for higher level care, and the likelihood of a severe episode was not ætiologically-related (Chi-squared test without trend, p = 0.987). There were 4 deaths in the study population: all were categorised as severe within organ dysfunction scoring systems.

Performance of APACHE II on admission

The distribution of admission APACHE II scores and numbers of patients subsequently utilising critical care is shown in Figure 1. The ability of APACHE II to predict the need for critical care support at various threshold scores is shown in Figure 2. The AUC for APACHE II at admission was 0.75 (95%CI: 0.66 to 0.84). At a threshold of 7, sensitivity is 72% and specificity is 66%, which is not adequately accurate to predict need for critical care. At this threshold, 28% of severe cases are missed (1-sensitivity) while a positive predictive value of 29% means that only one in three to four patients selected for critical care (of for study inclusion) would subsequently be classed as severe.

Performance of LODS, MODS, SOFA and APACHE II scores at 24 hours

Four markers commonly used in critical-care management were evaluated at 24 hours from admission and are shown in Figure 3. There is no one dominant test for all levels of test performance. If high specificity is required (>90%) then MODS, APACHE II and SOFA scores all perform similarly, whereas balanced sensitivity and specificity appears best achieved by the LODS measures. However, the overall AUC measures are similar for all four, with no statistically significant differences.

Performance of LODS, MODS , APACHE II, SOFA and CRP at 48 hours

At 48 hours the three organ dysfunction scores (LODS, MODS and SOFA) appear to perform better than APACHE II (a marker acute physiological derangement) and CRP (a marker of inflammation and necrosis): findings consistent with developing organ damage (Figure 4). Notwithstanding this qualitative finding, the overall area under curve measures are similar for all five, with no statistically significant differences. The SOFA score at cut-off of 1 achieved 79% sensitivity and 83% specificity, but this still only equates to a positive predictive value of 47%.

Identifying high and low risk patients

Test performance by test threshold value at 24 hours from admission is shown in Table 2. When looking to identify 'severe' cases (for recruitment to studies or target care) then sensitivity becomes a measure of selection efficiency and positive predictive value becomes a measure of selection accuracy. From Table 2, if a LODS score of one or more is used as the selection criterion then 90% of all (subsequently) severe cases will be included in those selected (sensitivity=efficiency) but only 35% of all recruits will (subsequently) be classed as severe (PPV=selection accuracy). The highest selection accuracy possible with LODS is only 50% and then only 7% of severe cases are being included. Looked at in this way none of the 24 hour test measures can achieve a recruitment purity of greater than about one half and so none are useful in isolation as a selection criterion to achieve a homogeneous risk group of patients.

The approach can also be used to identify 'mild' cases. In this instance specificity becomes the efficiency measure and negative predictive value (NPV) becomes the selection accuracy measure. For example if an APACHE score of 10 is taken as the threshold (scores of 9 or lower being selected) then selection accuracy will be 90% and efficiency will be 92%.

Similarly, at admission, APACHE II is an inadequate measure to predict severe cases. However, it performs quite well at detecting mild cases. A test threshold of 9 (scores of 8 of

less being selected) gives selection accuracy of 90% and efficiency of 70%. (This finding can be verified visually in Figure 1).

Combination scores

When individual tests fail to discriminate between different cases it may be possible to use a sequence of tests which together provide an adequate level of diagnostic certainty. The useful of this approach depends upon the tests providing sufficiently independent information or there is little or no increase in diagnostic yield. At 24 hours, APACHE II, LODS, MODS and SOFA scores are all moderately or highly correlated (Pearson's *r* ranging from 0.48 to 0.78, $p<0.01$ in all comparisons). A similar pattern was apparent for tests at 48 hours (APACHE II, CRP, LODS, MODS and SOFA scores: Pearson's *r* ranging from 0.48 to 0.67, $p<0.01$ in all comparisons). Consequently with the tests available, combinations are unlikely to be helpful in identifying severe and mild cases. In such instances, sensitivity and specificity can be traded against one another but it is not possible to improve both simultaneously.

Discussion

Although acute pancreatitis comprises a continuous spectrum of disease, the original concept within the 1992 Atlanta classification remains useful: the stratification of patients into mild and severe. Inadequate consideration of the definition of transient organ damage of other complications such as short-lived hypoxia or hypotension has led to coding inconsistencies between centres and over time, limiting the value of comparative studies which have used the Atlanta classification[16].

Accurately predicting severe and mild cases at (or soon after) disease onset might constitute a considerable advance in patient management and an important step towards reducing morbidity and mortality. In this context it is suggested that transient complications should not be coded as severe since these are (by definition) self-correcting with due diligence on the part of clinician teams.

Since 1992, major developments in critical care medicine have led to a number of validated organ dysfunction scores being developed, which have the advantage of being objective and reproducible. Findings presented show the inability of available tests to fulfil key roles in managing and researching severe acute pancreatitis. It would be beneficial to accurately and efficiently select severe patients to target the use critical care facilities, but this aim was not informed adequately by available tests. Additionally, it would be useful to accurately identify risk groups for inclusion in studies. In this instance efficiency of selection is more negotiable since the object is to recruit a homogeneous patient group with respect to severity. None of the test could adequately identify high risk groups although APACHE II scores could effectively select a low severity population. Consequently the selection of high risk patients for new clinical trials is fraught with difficulty, since selection guided by current tests will deliver a very mixed severity group.

Table 1: Test performance

		Disease	
		Severe	Mild
Test	+ve	a	b
	-ve	c	d

Intrinsic measures:
(read horizontally: the likelihood of positive or negative test finding being correct)

Positive Predictive Value (PPV) = $\frac{a}{a+b}$

Negative Predictive Value (NPV) = $\frac{d}{c+d}$

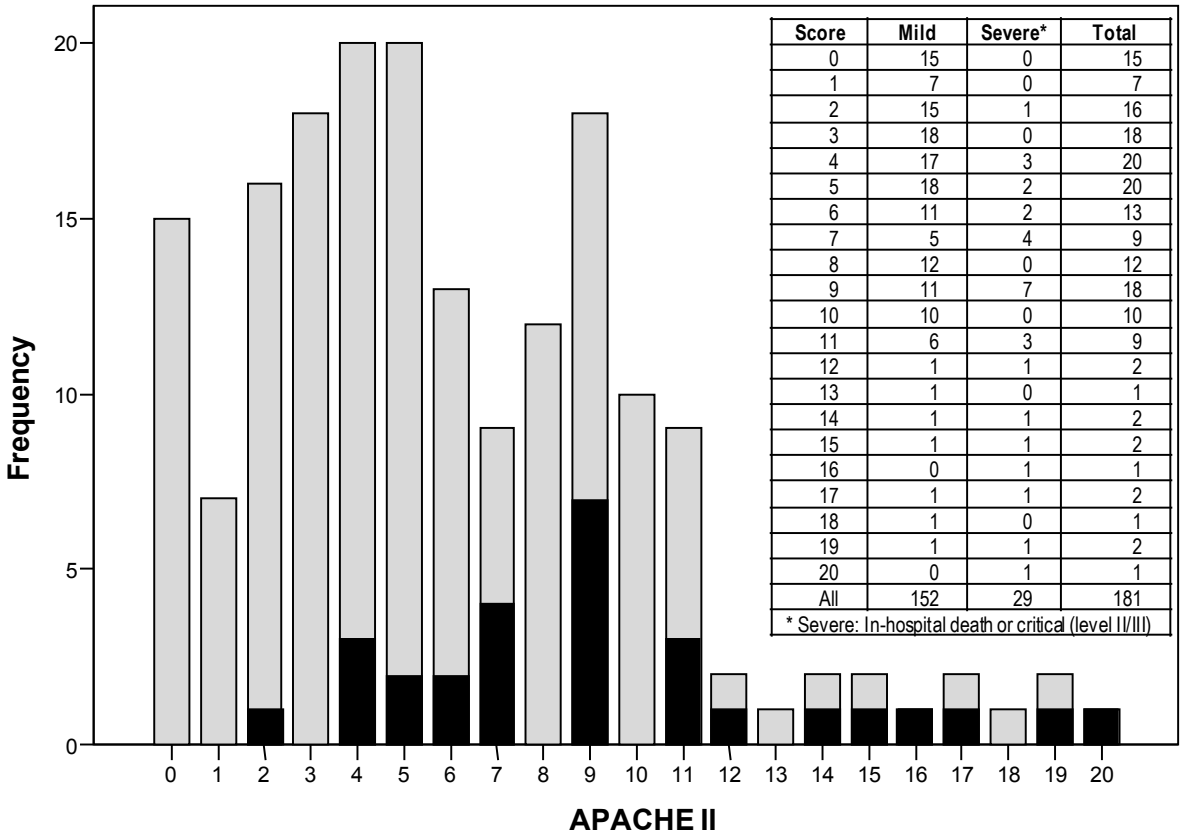
Extrinsic measures:
(read vertically: the likelihood that disease status is correctly identified)

Sensitivity = $\frac{a}{a+c}$, Specificity = $\frac{d}{b+d}$

Table 2: Acute pancreatitis test performance at 24 hours from admission by test threshold

Measure	Positive test threshold	Sensitivity	Specificity	Positive Predictive Value (PPV)	Negative Predictive Value (NPV)
LODS	1	90%	68%	35%	97%
	2	48%	85%	38%	90%
	3	28%	92%	40%	87%
	4	21%	95%	46%	86%
	5	14%	98%	57%	86%
	6	7%	99%	50%	85%
MODS	1	79%	72%	35%	95%
	2	45%	92%	52%	90%
	3	21%	97%	55%	86%
	4	10%	98%	50%	85%
SOFA	1	82%	51%	24%	94%
	2	64%	84%	44%	93%
	3	43%	95%	60%	90%
	4	14%	97%	50%	86%
APACHE II	2	100%	27%	21%	100%
	3	93%	45%	24%	97%
	4	90%	57%	28%	97%
	5	83%	62%	29%	95%
	6	66%	67%	28%	91%
	7	62%	76%	33%	91%
	8	62%	80%	38%	92%
	9	48%	87%	41%	90%
	10	45%	92%	52%	90%
	11	28%	95%	50%	87%
	12	17%	96%	45%	86%
	13	14%	97%	44%	85%
	14	14%	97%	50%	86%
	15	14%	98%	57%	86%

Figure 1: Distribution of APACHE II on admission and subsequent death of use of critical care (level II/III)



Score	Mild	Severe*	Total
0	15	0	15
1	7	0	7
2	15	1	16
3	18	0	18
4	17	3	20
5	18	2	20
6	11	2	13
7	5	4	9
8	12	0	12
9	11	7	18
10	10	0	10
11	6	3	9
12	1	1	2
13	1	0	1
14	1	1	2
15	1	1	2
16	0	1	1
17	1	1	2
18	1	0	1
19	1	1	2
20	0	1	1
All	152	29	181

* Severe: In-hospital death or critical (level II/III)

Figure 2: Predictive value of APACHE II on admission by cut-off threshold.
(key values tabulated)

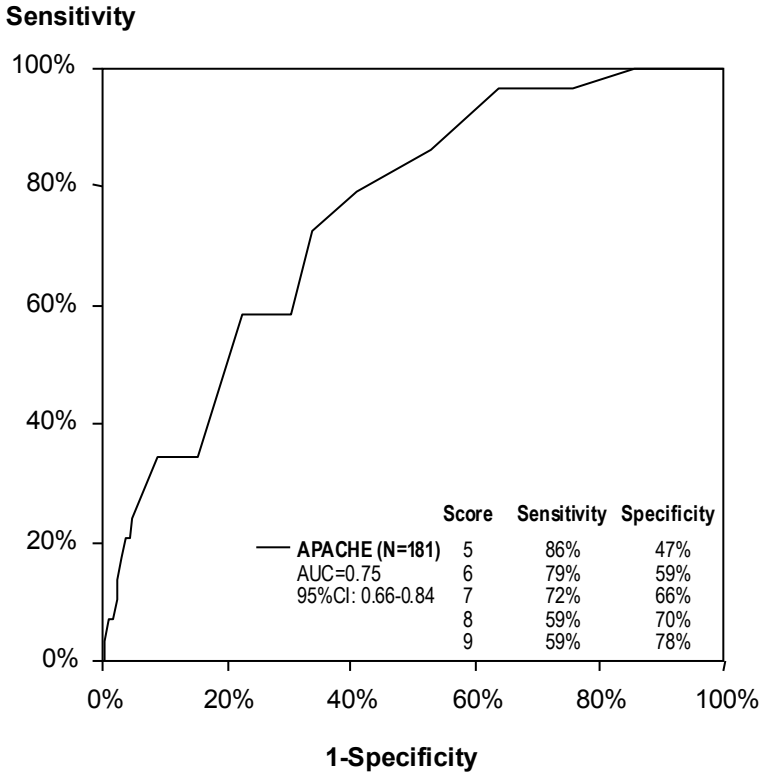


Figure 3: Critical care markers measured at 24 hours from admission

(key values tabulated)

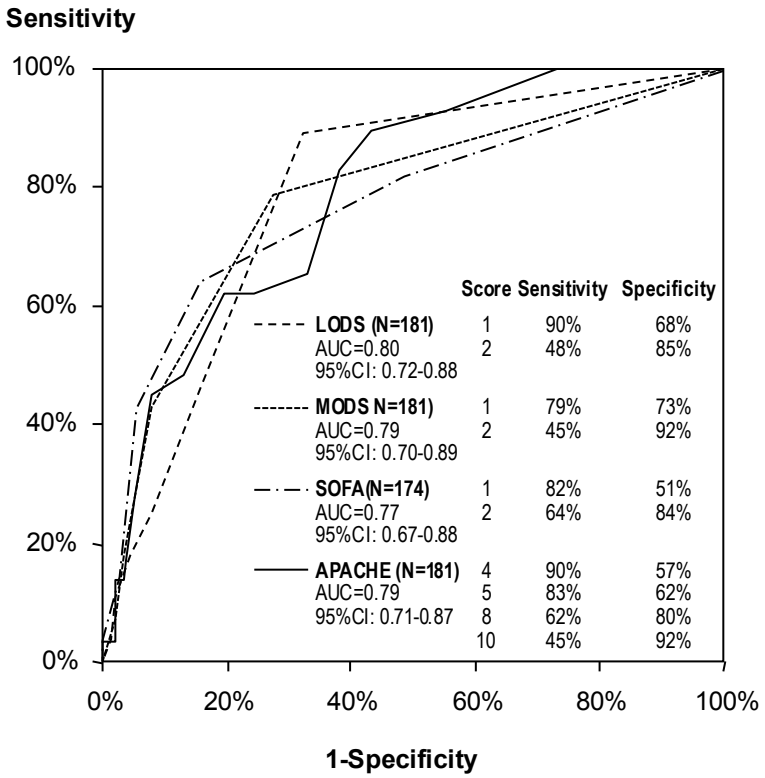
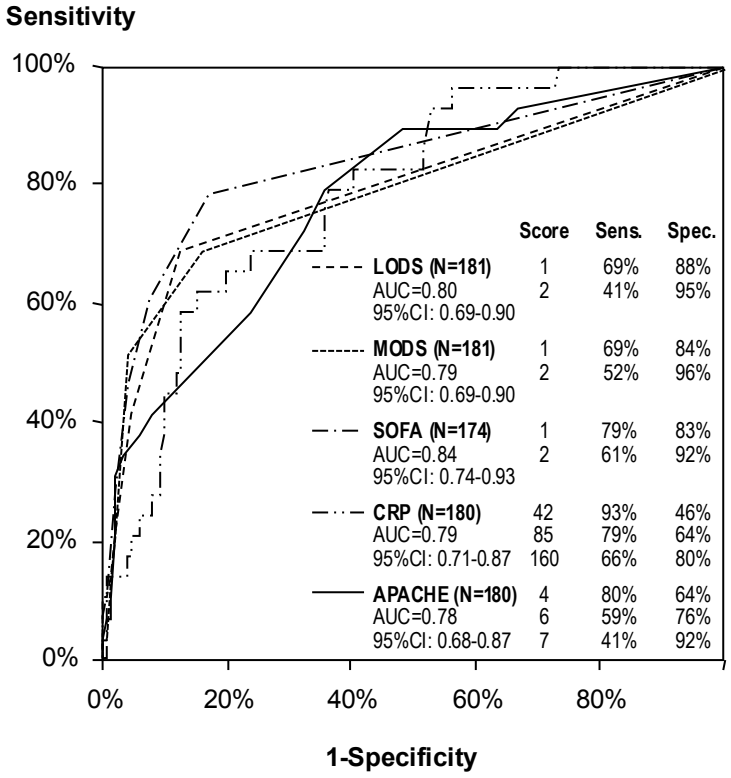


Figure 4: Critical care markers measured at 48 hours from admission

(key values tabulated)



References

1. Bradley EL, 3rd. A clinically based classification system for acute pancreatitis. Summary of the International Symposium on Acute Pancreatitis, Atlanta, Ga, September 11 through 13, 1992. *Arch Surg* 1993;128:586-590.
2. Powell JJ, Siriwardena AK. Pancreatobiliary emergencies. In: Garden O, ed. A *Companion to Specialist Surgical Practice.* 2nd ed. London: W B Saunders; 2001: 355-368.
3. Buter A, Imrie CW, Carter CR, et al. Dynamic nature of early organ dysfunction determines outcome in acute pancreatitis. *Br J Surg* 2002;89:298-302.
4. Mason J, Siriwardena AK. Designing future clinical trials in acute pancreatitis. *Pancreatology* 2005;5:113-115.
5. Halonen KI, Leppaniemi AK, Lundin JE, et al. Predicting fatal outcome in the early phase of severe acute pancreatitis by using novel prognostic models. *Pancreatology* 2003;3:309-315.
6. Mayer JM, Raraty M, Slavin J, et al. Severe acute pancreatitis is related to increased early urinary levels of the activation Peptide of pancreatic phospholipase A(2). *Pancreatology* 2002;2:535-542
7. Werner J, Hartwig W, Uhl W, Muller C, Buchler MW. Useful markers for predicting severity and monitoring progression of acute pancreatitis. *Pancreatology* 2003;3:115-127.
8. Larvin M, McMahon MJ. APACHE-II score for assessment and monitoring of acute pancreatitis. *Lancet* 1989;2:201-215.
9. Johnson CD, Toh SK, Campbell MJ. Combination of APACHE-II Score and an Obesity Score (APACHE-O) for the Prediction of Severe Acute Pancreatitis. *Pancreatology* 2004;4:1-6.
10. Powell JJ, Fearon K, Siriwardena AK. Current concepts of the pathophysiology and treatment of severe acute pancreatitis. *Br J Intensive Care* 2000;10:51-59.

11. Le Gall JR, Klar J, Lemeshow S, et al. The Logistic Organ Dysfunction system. A new way to assess organ dysfunction in the intensive care unit. ICU Scoring Group. *JAMA* 1996;276:802-810.
12. Marshall JC, Cook DJ, Christou NV, et al. Multiple organ dysfunction score: a reliable descriptor of a complex clinical outcome. *Crit Care Med* 1995;23:1638-1652.
13. Dervenis C, Johnson CD, Bassi C, et al. Diagnosis, objective assessment of severity, and management of acute pancreatitis. Santorini consensus conference. *Int J Pancreatol* 1999;25:195-210.
14. Department of Health U. Comprehensive Critical Care. A Review of adult critical care services. 2000:1-31.
15. King NK, Siriwardena AK. European survey of surgical strategies for the management of severe acute pancreatitis. *Am J Gastroenterol* 2004;99:719-728.
16. Bollen TL, van santvoort HC, Besselink MG et al. The Atlanta classification of acute pancreatitis revisited. *Br J Surg* 2008;95:6-21.