Update in Anaesthesia

Evidence-based medicine in critical care

Mark Davidson
Correspondance Email: mejdavidson@doctors.org.uk

Summary

Evidence based medicine (EBM) has an established role in management of critically ill patients. However ICU patients form a highly heterogeneous group and present challenges to the conduct of high quality randomised controlled trials. There are relatively few ICU studies that have stood the test of time, with the findings of the majority refuted and then disproved by subsequent larger studies or meta-analyses. This article describes the steps you should take when reading a study report, in order to decide whether it has been rigorously conducted, whether the results are valid and whether they can be applied to the patients that you see on a daily basis.

INTRODUCTION

Evidence-based medicine is an approach to medical practice in which the clinician integrates the best available clinical evidence from systematic research, with his or her own clinical expertise, to treat the individual patient in front of him. EBM and clinical expertise are not mutually exclusive - without reference to current best evidence, clinical practice will become out of date, while the adoption of an entirely evidence-based approach (if this is possible) risks exposing patients to inappropriate or inapplicable interventions. Conversely, treating patients solely on the basis of sensible and intuitive thinking risks exposing them to interventions which later prove to be harmful in subsequent clinical trials (e.g. intravenous beta agonists for Acute Respiratory Distress Syndrome, ARDS[2]), or denying patients beneficial therapies which were not previously widely practiced (low tidal volume ventilation for ARDS[3]). Clearly there is a middle ground to be found.

HIERARCHY OF EVIDENCE AND SOME PITFALLS

When assessing the evidence for a particular intervention there is now a widely accepted hierarchy of quality, with systematic reviews of published literature and meta-analyses seen as the most reliable form of evidence, and expert opinion as the least reliable. However not all randomised controlled trials (RCTs) are equal in quality and applicability.

Several recent high-profile cases of alleged research fraud[4-6] should alert readers to the fact that even a peer review process may not be sufficient to detect factitious or fraudulent research. Neither are the results of single RCTs necessarily reliable - the critical care literature is littered with examples of RCTs of interventions which showed early promise (activated protein C, tight glycaemic control, corticosteroids for septic shock), but where subsequent investigation has dampened early enthusiasm or in some cases entirely refuted initial claims. The dangers of changing practice based on evidence provided by a single trial or author are clear to see, and some experts recommend that two positive RCTs are needed to advise change of practice.

Not all interventions have a good evidence base behind them or necessarily require an evidence base to validate them – there has been no clinical trial of the use of parachutes for preventing death or injury from high altitude falls.[7] Die-hard advocates of EBM could point to the fact that there are examples of people falling from aeroplanes without parachutes and surviving, and falling with parachutes and dying, and that this illustrates the need for a properly conducted trial.

PROBLEMS WITH EBM IN CRITICAL CARE

As critical care grows as a specialty in its own right, there is an ever growing body of evidence available for the intensivist. However performing meaningful studies and then interpreting and applying the results presents particular difficulties in a critical care environment.

Critically ill patients are a heterogeneous group of patients, usually with multiple pathologies and comorbidities and at least one organ failure, who are subjected to variable and non-uniform treatments. Conducting a trial of therapy for ARDS, for example, a disease with a large number of precipitants that occurs in a wide spectrum of patients, means that the characteristics of the individuals within any sample

Box 1. *Hierarchy of evidence*

Systematic reviews and meta-analyses
RCTs
Cohort and case control studies
Case reports
Expert opinion

Increasing quality of evidence

Mark Davidson
Advanced Trainee in Intensive Care Medicine
Royal Devon and Exeter NHS Foundation Trust
Barrack Road
Exeter
Devon EX2 5DW
UK

group will vary greatly from each other. Although exposed to the same trial management protocols, the benefits or harms accrued may not be evenly distributed across the study population. This heterogeneity can only be accounted for in a study by using large sample sizes that give the trial adequate *power*. The size and complexity of large scale clinical trials makes them costly to run. Financial sponsorship is often provided from pharmaceutical companies, with an interest in subsequent marketing of the intervention, and conclusions of trials funded by for-profit organisations may be more likely to favour the study intervention[8]. This emphasises the importance of scrupulous attention to the conduct of the trial and interpretation of results.

Box 2. *Study power*

When designing a study the basic premise is taken that there is no difference between an intervention and the control (this is the *Null Hypothesis*).

In statistical testing two types of error may occur:

1. **Type 1 or alpha error** occurs when we reject the null hypothesis (that there is no difference between study groups) incorrectly i.e. our study finds a difference between groups where no difference really exists. Prior to data collection we decide on the value of alpha that we would find acceptable in this study - 0.05 (or 5%) is usually chosen. We will reject the null hypothesis if our P value is less than alpha. The chance that we will correctly accept a null hypothesis is (1 - alpha) is 0.95 (95%), and the chance we will incorrectly reject it is alpha, or 0.05 (5%).

2. **Type 2 or beta error** occurs when our study fails to find a difference between groups when, in reality, there is one. The power of a test is (1 – beta), and reflects the ability of a test to find a difference where there really is one.

The **power** of a test depends on a number of factors:

1. The *statistical significance* criteria used in the test (the more stringent the significance criteria, the more likely we are to accept the null hypothesis that there is no difference between study groups).

2. The *size of the treatment effect* in the population under study (larger effects will be detected, more reliably).

3. The size of the *sample population* (the larger the sample size, the more reliably a difference is detected, where one exists).

Power calculations should be performed prospectively (i.e. before the trial begins) to estimate the size of the sample population needed to find the difference in question. Power may be calculated retrospectively to calculate the beta error rate given the sample size.

While not limited to critical care, the problems associated with duplicate publishing of positive trial results or failure to publish negative trial results (forms of publication bias - see below), and delayed publishing of results (e.g. the Tracman study) mean that even the most thorough review of the literature may not provide the clinician with all the information which should be available to him to make the correct decision for a patient.

CRITICAL APPRAISAL OF A STUDY

In assessing the usefulness of a particular RCT we need to assess its internal and external validity. The study should be internally valid – that is to say that the results and conclusions drawn from it, by the authors, should be consistent with the design and conduct of the study itself. If we wish then to apply the study result to the particular clinical setting in which we work, we must also assess the study's external validity (also termed generalisability or applicability).

In the following section of this article we will appraise a recent RCT, examining mortality after a fluid bolus in African children with severe infection[1] (summarised in the box below). To do this we will try to answer three questions (using guidelines produced by the Centre for Evidence Based Medicine at the University of Oxford, UK):

1. Are the results internally valid?
2. What are the results?
3. Are the results externally valid?

Maitland K, Kiguli S, Opoka RO et al.

Mortality After Fluid Bolus In African Children With Severe Infection.

N Eng J Med 2011; **364**: 2843-95.

Available at http://www.nejm.org/doi/pdf/10.1056/

NEJMoa1101549

The FEAST (Fluid expansion and supportive therapy) trial was conducted at six hospitals in sub-Saharan Africa (4 in Uganda, 1 in Kenya and 1 in Tanzania).

The study enrolled children aged between 60 days to 12 years of age, suffering a febrile illness with either impaired consciousness, respiratory distress or both, and with impaired perfusion, indicated by one or more of the following:

- a capillary refill time of 3 or more seconds,
- a lower limb temperature gradient,
- a weak radial-pulse volume, or
- severe tachycardia.

Patients were randomly assigned to receive 20ml.kg^{-1} 0.9% saline or 20ml.kg^{-1} 5% albumin or no bolus of fluid. At 1 hour they were administered an additional 20ml.kg^{-1} 0.9% saline or 5% albumin, if they still had signs of poor perfusion. Children with infective gastroenteritis and severe malnutrition were among those excluded from the study. The primary endpoint was mortality at 48 hours and it was assumed that this would be about 15% in the control (no bolus) group.

All other therapies - maintenance fluids, antimalarials, antipyretics, anticonvulsants and transfusion parameters

(20ml.kg^{-1} if Hb less than 5.0g.dl^{-1}) - were the same for both treatment groups and the control group.

3141 children entered the main arm of the study; however the study was stopped at this stage (short of the planned 3600 patients), due to safety concerns in the intervention groups. The baseline characteristics of the three groups were similar. Both bolus groups received 20ml.kg^{-1} in the first hour of treatment, compared to 1.2ml.kg^{-1} in the non-bolus group, demonstrating that fluid protocols were followed accurately in all groups.

The mortality rates at 48 hours were 10.6% in the albumin bolus group, 10.5% in the saline bolus group and 7.3% in the no bolus group. Children were 1.45 times more likely to die if given a fluid bolus, compared to the children who received no fluid bolus (95% confidence intervals, 1.13-1.86; p=0.003). This difference in outcome was still clear at 4 week follow-up. There was no difference between the albumin and saline bolus groups. Further analysis shows that children who were not given fluid boluses had better survival in almost all subgroups, including those with malaria, profound anaemia, worse acidosis and higher lactate levels.

Are the results of the study internally valid?

We want to know whether the treatment effect reported in the article represent the true direction and magnitude of the treatment effect under investigation. In other words, do these results represent an unbiased estimate of the treatment effect, or have they been influenced in some systematic fashion leading to a false conclusion?

A bias is a systematic error in the results of an investigation, which may underestimate or overestimate the true effect of the intervention. Sources of bias are described in the table below.

There are a number of questions to help us answer this question about internal validity.

Was the assignment of patients to treatments randomised?

Randomisation was performed by a recognised and robust method, permuted blocks of random sizes (see below). It was performed at a remote location, stratified by centre, and used sequentially numbered, opaque, sealed envelopes. Clinicians would have had negligible influence on the allocation of patients to the study or control group, should they have had a preference. Permuted block randomisation ensures that, at any stage of the study, roughly equal numbers of participants have been allocated to each group. The use of random block sizes ensures that knowing the group allocations of the previous subjects doesn't confer an ability to guess the allocation of (and perhaps, therefore, influence the decision to enrol) the next subject.

Randomisation is used to minimise selection bias to ensure that treatment and control groups are as similar as possible, apart from the treatment under investigation. This is achieved by distributing covariates (i.e. variables which may be either independently predictive of outcome, or confounders) equally between study groups, which allows valid comparisons to be made between groups. The beauty of randomisation is that it should allow equal distribution of both known and unknown factors between groups.

Allocation concealment refers to the steps taken to prevent patients or clinicians knowing prior to randomisation, to which group the next patient will be allocated.

Were the groups similar at the start of the trial?

If the randomisation process has been successful there should be little difference in baseline characteristics between treatment and control groups. The characteristics reported should include recognised covariates or variables, which could affect study outcome. When the

Table 1. *Sources of bias*

Type of bias	Explanation	Avoidance of bias
Selection bias	Systematic differences between baseline characteristics of groups being compared.	Randomisation and allocation concealment should be carried out satisfactorily.
Performance bias	Systematic differences between groups in the treatment that is provided, or in exposure to factors other than the interventions of interest.	Groups should be treated identically except for the intervention in question.
Attrition bias	Systematic differences between groups in rates of withdrawal from a study.	Analysing results on an intention-to-treat basis reduces this.
Detection bias	Systematic differences between groups in how outcomes are determined.	Blinding participants and outcome assessors reduces this.
Reporting bias	Systematic differences between reported and unreported findings.	

study groups are small, chance may place those with apparently better prognosis or confounding variables, in one group, but as sample size increases this becomes less likely. Although frequently seen, statistical analysis (and therefore reporting 'p-values') to describe the random baseline differences between groups is meaningless – any differences observed have, by definition, occurred by chance.

In our example study, mid-upper arm circumference was measured as a surrogate marker of nutritional status (which could potentially have an effect on mortality in severe sepsis) and malnourished patients were distributed equally across groups. In fact there were no clinically significant differences between the study groups in any of the baseline characteristics, which increases our confidence that the randomisation process was effective.

Aside from the allocated treatment, were the groups treated equally?

The two study groups should be treated equally in all respects other than the treatment under investigation, to ensure that differences in outcomes between groups are due to the treatment under investigation, and not some other (known or unknown) factor.

In our example study this appears to be true. All patients received intravenous maintenance fluids, antibiotics, anti-malarials, antipyretics, anticonvulsants, treatment for hypoglycaemia and blood transfusions as appropriate. They were treated in the same hospitals and received the same follow up.

Were all patients who entered the trial accounted for and were they analysed in the groups to which they were randomised?

The greater the number of subjects who are lost to follow-up, the more the trial may be subject to attrition bias. Lost patients may not have outcomes similar to the group they were lost from (e.g. all lost patients may have died, or all may have recovered to the extent that they did not return for reassessment). If few patients have the outcome of interest, then even small losses to follow up can bias the trial result. Loss of subjects should be minimal, preferably less than 20%.

The principle of attributing all patients to the group to which they were randomised, results in an **intention-to-treat analysis** (even if they received no treatment or crossed-over and received treatment in the other arm of the study). It preserves the value of randomisation and minimises some sources of bias.

Also of interest is the total number of patients assessed for eligibility for entry into a trial and the reasons for exclusion of those deemed ineligible. Low rates of recruitment into RCTs are known to be due, in part, to additional (non-specified) selection by participating clinicians. Furthermore it is known that patients recruited into RCTs can differ from those eligible, but not recruited in terms of age, sex, race, disease severity.

The example study provides a comprehensive analysis of the flow of patients from assessment of eligibility to statistical analysis. All patients were analysed on an intention-to-treat basis. Losses to follow-up were approximately 2.5% and equally distributed among the study groups.

Were measures objective, or were clinicians kept 'blind' to the treatment received?

Blinding occurs when either the patient, clinician, or both ('double-blind') are unaware of the group to which the subject has been allocated. Preconceived opinions about a treatment, whether pessimistic or optimistic, can systematically bias other aspects of treatment and the reporting of treatment outcomes.

Blinding reduces the risk that it was the knowledge of which intervention the subject received, rather than the intervention itself, that affected outcome. This is more important for subjective outcome measures (e.g. scores of symptom relief or functional improvement) than objective measures (e.g. death or stroke). Blinding may not always be possible (e.g. in the case of surgery) but is desirable where practicable. In the SAFE study of 0.9% saline versus albumin for resuscitation of the critically ill, both treatment fluids were delivered in glass bottles, concealed within a cardboard box, in order to blind clinicians.

In this study neither the clinician administering, nor the patient receiving, a fluid bolus could be blinded. However, an assessment of neurological sequelae four weeks after recruitment was performed by an assessor who was unaware of (or blinded to) the treatment assignments. Therefore blinding has been performed where possible and applied to the most subjective outcome measure.

What were the results?

Dichotomous results ('yes' or 'no' outcomes such as death or myocardial infarction) are usually used where possible, as statistical analysis is more straightforward and the results more meaningful.

How large was the treatment effect?

A full discussion of statistical tests is beyond the scope of this article, but we will review some of the basic outcome measures that are commonly used (see Box 2).

Consider an intervention which is designed to reduce mortality in patients with severe sepsis. The relative risk (RR) tells us how many times more likely it is that an event will occur in the treatment or experimental group, than the control group. A relative risk of 1 means that there is no difference in the outcome measure between the two groups. If the RR is less than one, the outcome is less likely in the treatment group, and conversely a RR > 1 means it is more likely.

Absolute risk reduction (ARR) tells us in absolute terms the difference in risk (or rates) of the outcome between treatment and control groups. An ARR of zero means the outcome is equally likely in treatment and control groups.

Relative risk reduction (RRR) tells us the reduction in risk of the outcome relative to the risk of the outcome in the control group. The control event rate (CER) is important here - consider an intervention which has a RRR of 30%. If the risk of death in the control group is 10% (i.e. CER = 0.1), then a RRR of 30% reduces the risk of death to 7% over whatever period the intervention and study were applied. However, if the CER for the intervention was low (say 0.1%) the

Box 3. *Demonstration of outcome of patients in a RCT for control and intervention groups.*

	Dead	Alive	Total
Treatment	a	b	a + b
Control	c	d	c + d

Control event rate (CER) = c/(c+d)

Experimental event rate (EER) = a/(a+b)

Relative risk (RR) = EER/CER

Absolute risk reduction (ARR) = CER – EER

Relative risk reduction (RRR) = ARR/CER

Number needed to treat (NNT) = 1/ARR

same RRR of 30% would represent a much less meaningful benefit to patients, particularly when weighed against the cost of the intervention and the risk of adverse effects. The RRR is the most often reported outcome measure, perhaps because it provides a numerically larger estimate of treatment effect than the ARR, when the CER is low.

The *number needed to treat* (NNT) is arguably the most clinically relevant measure of outcome and represents the number of patients we need to treat with the experimental intervention, to prevent one adverse outcome. When the intervention causes more harm than the control the NNT is negative and is usually converted to a positive number and expressed a *number needed to harm* (NNH).

In the study of children who received saline boluses versus no boluses, the RR of death was 1.44 (i.e. death more likely in the saline group). In other words, children were 44% more likely to die in the saline group compared to the control group. This is a large effect given a control event rate of 7%. ARR = (0.073 – 0.106) = -0.033, and the NNH = 30 (1/0.033).

How precise was the treatment effect?

By convention we consider a study to be positive (i.e. showing a difference between the intervention and the control) if the statistical analysis shows that we are 95% sure that the study result represents a true difference between the intervention and the control. Put another way, the study result will not be a true representation of 'the truth' 5% of the time - if 20 identical studies were conducted, 19 would agree and give this result, but one of them would show the opposite result.

The true relative risk of an intervention, applied to an entire population, can never be known, but a rigorous controlled trial can provide an estimate of the treatment effect in a sample of the population (the trial subjects) at a given point in time - **a point estimate**. The true value of the RR of the treatment lies somewhere in the range defined by the study; confidence intervals are used to describe this range. Ninety-five percent confidence intervals of the RR tell us that we can be 95% sure that the true treatment effect (or RR) is within the range quoted.

If the confidence interval is narrow the point estimate of the population RR is an accurate reflection of the true population value (provided the results are not subject to significant bias). If the 95% confidence interval, overlaps a RR of 1.0, i.e. the value corresponding to no effect, then the results are not statistically significant. If the value corresponding to no effect (RR = 1) lies outside the 95% confidence interval, then the results are statistically significant at the 5% level (i.e. the result could occur by chance less than 5% of the time).

The calculation of confidence intervals for relative risks is complicated and beyond the scope of this article.

In our study of children who received saline boluses compared to no boluses, the RR of death was 1.44 with 95% confidence intervals of 1.09-1.90. This confidence interval does not include unity (1.0), so the results for this comparison are significant at the 5% level. A similar result was true for the comparison of albumin boluses versus no boluses.

Will the results of the study help me in caring for my patient (or are the results externally valid)?

We have established that the study demonstrates internal validity - it has been conducted rigorously and the results are probably a true representation of this population. We should now consider whether it is applicable to other patients in other clinical settings (its external validity). An excellent article by Rothwell highlights some of the pitfalls surrounding this problem.[9] There are a number of questions we should ask ourselves, before deciding to apply the results of a study to patients in our care.

Are my patients so different from those in the study that the study results do not apply?

Ideally, we want to ask ourselves whether the patient in front of us would have met the inclusion criteria for the study, and not fulfilled the exclusion criteria; the answer is rarely a straightforward 'yes'. By their nature RCTs study treatment effects in the context of a clinical trial, and not in general clinical practice, so inevitably external validity will be less than perfect. In reality, a treatment effect will be influenced by factors such as the doctor-patient interaction, the placebo effect, doctor or patient preference etc. All of these factors are minimized in clinical trials by the use of blinded allocation of treatments, placebo

control, and the exclusion of clinicians who do not have equipoise over the intervention in question (i.e. excluding clinicians who hold a prior belief that one or other of the study interventions is superior). The net effect of these factors probably underestimates the benefits of treatment in clinical practice.

The **placebo effect** is the name given to the benefits or changes in outcome measures perceived by patients or assessors, when an inert treatment is administered. Treatments are compared with an appropriate control to ensure findings are not due to this placebo effect alone. It is more important when subjective outcome measures are being used (scores of pain, satisfaction, quality of life etc) than objective measures (mortality, heart rate etc).

The setting in which the trial was performed is clearly important. Differences between healthcare systems (and even between countries, which operate similar healthcare systems) have been shown to affect external validity. For example, trials testing the BCG vaccination for the prevention of tuberculosis, demonstrated great effectiveness in more northern countries, with far less effect in trials conducted further south. Additionally, there may be significant differences in the use of ancillary non-trial treatments – a particular treatment may be considered standard practice in one country, for a particular condition, but it may be rarely used elsewhere.

Selection of centres to conduct clinical trials has the potential to affect external validity. While it may be tempting to perform a trial in specialist intensive care units of a large metropolitan teaching hospitals, the trial results may be more generalisable if a wider variety of hospitals are included in the trial. Consider a trial of glycaemic control in a Belgian intensive care unit[10] in which over 60% of the patients were post cardiac surgery. This study found a relative risk reduction in ICU mortality of 42%, in patients randomised to tight glycaemic control (blood glucose levels 4.4-6.1mmol.L^{-1} or 80-110mg.dL^{-1}) compared to conventional treatment (blood glucose levels less than 11.1mmol.L^{-1} or 200mg.dL^{-1}), findings which were not replicated in a medical ICU by the same author[11] or a large international multicentre randomised controlled trial by the ANZICS study group.[12]

Many studies exclude pregnant women, the young and the elderly for ethical or other reasons, so care must be taken when extrapolating a trial's results to these populations.

The use of 'run-in' periods can be more difficult to recognise. Patients in a placebo run-in all receive placebo to assess patient compliance with trial protocols, and those patients who are poorly compliant are excluded from analysis. Excluded patients are known to differ from recruited patients in age, social class etc. Active treatment run-ins, in which patients who are intolerant of the study intervention are excluded, produce trial data with much lower complication and treatment failure rates than may otherwise be seen, and can seriously undermine external validity.

Some clinical trials use enrichment strategies. In these trials patients are selected who are likely to respond well to treatment, or perhaps even had a previous good response to a similar drug. Although there may be a place for such trials, their external validity is clearly much reduced.

What of our study? Over three thousand patients were randomised over two years across six centres (~0.6 patients/centre/day). The study was performed in resource-poor healthcare systems, where over half of the children presenting with severe infections had P. falciparum parasitaemia. Those with bacterial sepsis (only 12% had a positive blood culture) may have benefited from fluid therapy, however there are logical reasons why those with pneumonia, cerebral malaria and other causes of encephalopathy may have been harmed by fluid therapy, as they have high levels of ADH (antidiuretic hormone) resulting from their underlying disease.

In addition, the high numbers of children with severe anaemia may be harmed by liberal fluid administration, since haemodilution in profoundly anaemic children may reduce oxygen delivery below a critical level needed for adequate organ oxygenation. The mortality rate in the control group (7.3%) was considerably lower than predicted (15%) and it is likely that this reflects the training in triage, basic life support and regular monitoring that was introduced as part of the study.

A further limitation of the generalisability of this study is that children with dehydration due to gastroenteritis were excluded, and it would be disastrous if fluid therapy were withheld from children with this form of septic illness.

Were all clinically important outcomes considered?

Surrogate outcomes are used as indirect measures of clinical outcome, but the literature contains many examples of studies in which treatments had apparently beneficial effects on surrogate markers of outcome, but subsequent RCTs, with appropriate clinical endpoints, showed the treatments to be either ineffective or harmful. A good example is the CAST trial[13] in which anti-arrhythmic agents, such as flecainide, were administered to patients after myocardial infarction on the basis that they reduced ECG abnormalities in pilot studies. Mortality was increased in the treatment arm of this RCT.

Outcome measures should be patient-centered (i.e. provide information the patient wants), avoid the use of composite outcomes (e.g. "stroke or cardiac death" as outcomes may vary between components of the composite outcome) where possible, be measured after adequate length of follow up, and report adverse effects of treatment.

In our study, the primary (dichotomous and meaningful) endpoint was death at 48 hours. Secondary endpoints concerning adverse effects of volume overload were also considered, for example there appeared to be no statistically significant increase in pulmonary or cerebral oedema, two features that might have explained the excessive mortality caused by fluid boluses.

CONCLUSION

This study of fluid boluses in the resuscitation of African children with severe infections appears to be methodologically sound, and has found increased mortality in the treatment arm which is statistically (and clinically) significant. It was conducted in a region with developing economies and healthcare systems in an area with a high prevalence of malaria due to P. falciparum. Its generalisability is limited to patients and clinicians working in similar circumstances.

Evidence based medicine is a powerful tool, which has a clear place in modern medical practice. Obsessive adherence to EBM risks taking the art out of medicine, to the detriment of our patients care. In contrast, refusal to accept and apply the results of good quality RCTs, on the basis of insufficient external validity, is not acceptable practice either. For the foreseeable future, we must rely on EBM and an enquiring mind, tempered by some healthy skepticism and empathy with our patients, to help us find fitting strategies to guide our patients though their critical care admissions.

REFERENCES

1. Maitland K, Kiguli S, Opoka R et al. Mortality After Fluid Bolus In African Children With Severe Infection. *N Eng J Med* 2011; **364**: 2843-95. Available at: http://www.nejm.org/doi/pdf/10.1056/NEJMoa1101549
2. Smith FG, Perkins GD, Gates S et al., Effect of intravenous β-2 agonist treatment on clinical outcomes in acute respiratory distress syndrome (BALTI-2): a multicentre, randomised controlled trial. *Lancet* 2012; **379**: 229-35.
3. ARDSNET Investigators. Ventilation with lower tidal volumes as compared with traditional tidal volumes for acute lung injury and the acute respiratory distress syndrome. *N Engl J Med* 2000; **342**: 1301-8.
4. Shafer SL. Editor's Note: Notices of Retraction. *Anesthesia and Analgesia* 2011; **112**: 1246-7.
5. Shafer SL. Notice of retraction. *Anesthesia and Analgesia* 2009; **108**: 1350
6. Roberts I, Smith R, Evans S. Doubts over head injury studies. *BMJ* 2007; **334**: 392-4.
7. Smith GCS, Pell JP. Parachute Use To Prevent Death And Major Trauma Related To Gravitational Challenge: Systematic Review Of Randomised Controlled Trials. *BMJ* 2003; **327**: 1459-61
8. Als-Nielsen B, Wendong C, Gluud C et al. Association of Funding and Conclusions in Randomized Drug Trials A Reflection of Treatment Effect or Adverse Events? *JAMA* 2003; **290**: 921-8.
9. Rothwell PM. Treating Individuals 1: External Validity Of Randomized Controlled Trials: "to whom do the results of this trial apply?" *Lancet* 2005; **365**: 82-93.
10. Van den Berghe G, Wouters P, Weekers F et al. Intensive Insulin Therapy in Critically Ill Patients. *N Engl J Med* 2001; **345**: 1359-67.
11. Van den Berghe G, Wilmer A, Hermans G et al., Intensive insulin therapy in the medical ICU. *N Engl J Med* 2009. **360**: 1283-97.
12. NICE-SUGAR Study Investigators, Finfer S, Chittock DR, Su SY et al. Intensive versus conventional glucose control in critically ill patients. *N Engl J Med* 2006; **354**: 449-61.
13. The Cardiac Arrhythmia Suppression Trial (CAST) Investigators. Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. *N Engl J Med* 1989; **321**: 40.