RESEARCH METHODS & REPORTING

# Dangers of non-specific composite outcome measures in clinical trials

Composite outcomes seem an attractive method to increase statistical power, but they can mask the effect of treatment

David Prieto-Merino *lecturer*[1], Liam Smeeth *professor*[2], Tjeerd P van Staa *professor*[2 3], Ian Roberts *professor*[1]

[1]Clinical Trials Unit, London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK; [2]Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine; [3]Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Utrecht, Netherlands

According to international guidelines,[1 2] outcome measures in a clinical trial should address the risks and benefits of a treatment, be relevant to patients, and be sufficiently common to make the trial feasible. In an attempt to meet these objectives many investigators select outcomes such as all cause mortality, all hospital admissions, or any adverse event. These outcomes can be thought of as composite outcomes as they combine multiple outcomes that are cause specific. All cause mortality is a popular outcome measure because it is believed to provide the net effect of the treatment, it seems more patient relevant than cause specific mortality, and it provides more outcomes so should increase statistical power. Another common approach to increase power is to use wide case definitions and sensitive tests.

In recent years several papers have reviewed and debated the use of composite outcomes in clinical trials.[3-9] Authors agree on their advantages and disadvantages, and a good summary can be found in a recent report from the European Network for Health Technology Assessment.[10] Briefly, the main objection to the use of composite outcomes is that, if the treatment has different effects on the different components of the outcome, the net effect on the composite outcome is difficult to interpret. It also complicates patient management decisions, raising the question of which particular component outcome is more relevant for each patient.

However, in the ongoing debate little emphasis has been given to the fact that, by including events that are not causally related to the treatment (either by using a composite outcome or by misclassifying events with a wide case definition), the overall effect of the trial will be diluted towards the null. In this paper we explain why dilution occurs, provide examples of trials where this has happened, and discuss how dilution can offset many of the supposed advantages.

## Motivating example: planning a clinical trial

Having reviewed the existing evidence, you believe that β blockers might improve lung function and so reduce hospital admissions in patients with chronic obstructive pulmonary disease (COPD), although you are worried that they might cause severe bronchospasm in some patients. You decide to conduct a randomised controlled trial comparing β blockers with placebo in people with moderate COPD. Because of the large burden on health services and the importance for the patients, you define the primary outcome as the risk of hospital admission over one year. Your practice records show that patients with moderate COPD have a 40% risk of being admitted to hospital each year (although only half of admissions are for COPD). The statistician says that a study of about 1510 patients followed up for a year should have enough power to detect a 20% decrease in admissions (that is, from 40% to 32%). Because of your concern about bronchospasm, participants are asked to keep a diary of episodes of wheeze or shortness of breath as well as more severe symptoms. Severe bronchospasm is rare in people with COPD (about one episode per 10 person years), although many feel wheezy or short of breath (about once a week).

Imagine that trial recruitment went well and that table 1⇓ shows the (hypothetical) results. There is no significant reduction in hospital admissions and no evidence of increase in the adverse event. Should you conclude that β blockers have no beneficial or adverse effects?

This example has been made up to show how real treatment effects can be diluted by using non-specific outcome measures. Patients are admitted to hospital for many different reasons. “Any hospital admission” can be considered a composite of different cause specific admissions. In this example, half of admissions were related to COPD. If we examine only COPD related admissions there is some evidence of a treatment effect.

Correspondence to: D Prieto-Merino david.prieto@lshtm.ac.uk

We see the same thing with adverse events. Even though β blockers did increase the risk of severe bronchospasm, the ability of the trial to detect this effect is lost altogether because the outcome measure includes wheeze and shortness of breath, which are common in COPD patients and are not affected by the intervention. If outcomes that are causally related to the trial treatment are combined with those that are not, the estimate of the treatment effect is diluted towards the null and we may fail to identify potentially important benefits or harms.

## Explaining treatment effects with a causal mechanism model

The effect of a treatment on a health outcome can be shown using the causal mechanism model described by Rothman and colleagues.[11] For a patient to experience a specific health outcome, a particular set of events, conditions, or characteristics must occur. This set is called the causal mechanism. The causal mechanism is considered sufficient to produce the health outcome, and each component cause is necessary in the sense that had it been absent, that outcome would not have occurred. The same health outcome might result from different causal mechanisms sharing some component causes, and different outcomes might arise from similar causal mechanisms. A trial treatment has the potential to cause a change in health outcome by blocking (or adding) one or more of the necessary component causes in the causal mechanism. We would say that a health outcome is causally related to a treatment if some causal mechanisms contain component causes that are blocked by the treatment. Figure 1⇓ shows a hypothetical trial with three outcomes: red, blue, and green. For simplicity, we assume that the same number of patients was allocated to the treated and untreated groups, so that the relative risk is the ratio of patient outcomes. Treatment can prevent some (but not all) of the red outcome. In this example, the red outcome is causally related to the treatment whereas the blue and green outcomes are not.

## Composite outcomes

A composite outcome combines different outcomes (with different causal mechanisms).[1] The effect of a treatment on a composite outcome is a weighted average of its effects on the outcomes that are combined. In fig 1 the treatment halves the occurrence of the red outcome (relative risk =0.5) but has no effect on the blue outcome (relative risk=1.0). The relative risk of the outcome red or blue is the weighted average of the relative risks for the red and blue outcomes: relative risk red or blue=0.5(10/13)+1.0(3/13). Where (10/13) and (3/13) are the relative weights of the red and the blue outcomes in the untreated group. If a composite includes outcomes that are not causally related to the treatment, the relative risk for the composite is diluted towards the null.[12] Because few treatments will be causally related to all causes of death, all cause mortality is a composite outcome that combines causally related causes of death with those unrelated to the treatment.

## Misclassification of outcome

Many trials use diagnostic tests to assess the presence or absence of a health outcome. For example, troponin levels might be used to determine the presence or absence of a myocardial infarction.[13] Whatever method is used to assess outcome, there will be false positives and false negative results. The number of false positives divided by the total number of positive outcomes is the proportion of false positives. Figure 2⇓ shows the effect of outcome misclassification on the relative risk in a trial. Each arm is assumed to have 2000 participants and each circle represents 20 outcomes. The treatment effect estimated using the causally related outcome is 100/200 (relative risk=0.50, 95% confidence interval 0.40 to 0.63). Some related outcomes have low biomarker values and some unrelated outcomes have high biomarker values. The misclassification of outcomes, the proportion of false positives, and the relative risk will change depending on the biomarker's cut-off value. If we define the outcome as present when the biomarker value is more than seven, we will miss 20 causally related outcomes in the treatment group and 40 in the untreated group. We also add 20 false positive outcomes to both groups. Because false positive outcomes are included, the relative risk moves closer to the null (relative risk=0.56, 0.44 to 0.70). The dilution is greater if we define the outcome as present when the biomarker value is more than three, because there are even more false positive results (0.62, 0.51 to 0.74). Rodgers and McMahon[14] showed that, whatever the true relative risk, the estimated relative risk moves towards the null as the proportion of false positive results increases.

Figure 2⇓ also shows that as the number of outcomes increases, the precision of the relative risk increases (narrower confidence interval). However, in this case we are more certain about the answer to a different question. With the biomarker cut-off at 3, we have a precise estimate (0.62, 0.51 to 0.74) but one that excludes the true relative risk of 0.5. Notice that, on the other hand, the higher the value required in the biomarker, the more real outcomes that we will fail to detect (false negative results). The increase in false negative results will reduce the power of the analysis but will not cause dilution as this is only caused by the inclusion of false positives.

## Examples from real trials

The CRASH-2 trial is a randomised placebo controlled trial of tranexamic acid versus placebo in bleeding trauma patients.[15] The primary outcome was "all-cause mortality in hospital within 4 weeks of injury." However, tranexamic acid was expected to work by reducing bleeding in which case death from bleeding is arguably the causally related outcome. The relative risk for death from bleeding with tranexamic acid was 0.85 whereas the relative risk for non-bleeding deaths was 0.94. In the untreated group, 36% of deaths were due to bleeding and 64% to other causes. The relative risk on all cause mortality was the weighted average of the two effects: relative risk=0.91=0.85×0.36 + 0.94×0.64 (fig 3(*a*)⇓).

The Heart Protection Study is a randomised controlled trial of simvastatin versus placebo in 20 500 patients at high risk of cardiovascular events.[16] The aim was to evaluate effects of cholesterol lowering on patient outcomes. There were concerns about a possible increase in the risk of myopathy, defined as "unexplained muscle pains or weakness" (myalgia) with raised levels of the enzyme creatine kinase. Reports of myalgia were collected from trial participants and creatine kinase was measured in these patients. Figure 3(*b*) shows the effect of simvastatin on myalgia for different cut-off values of creatine kinase. As the cut- off value is reduced there are more myalgia outcomes and so the confidence interval is narrower. However, the relative risk moves closer to the null. Myalgia is a common symptom and is likely to have many different mechanisms. Most cases will not be causally related to the treatment. Nevertheless, the possibility that some cases might be related to treatment cannot be excluded from these data.

## Discussion

Composite outcomes are often used to obtain more outcome events and thus increase statistical power.[17] This is reasonable if all components of the composite are causally relevant and the treatment effects are similar. If the treatment effects vary considerably between outcome components the trial provides a more precise answer but to an irrelevant average effect that does not represent any meaningful clinical effect. Furthermore, if some components are not causally related to the treatment, the effect of treatment will be diluted towards the null, offsetting the gain in precision and not increasing the power of the trial (as shown in the examples above). The real solution may be to increase the sample size until there is enough power to answer the question for the outcome of interest.

The CONSORT guideline recommends that the main outcome should be something "relevant to the patient."[2] It is tempting to assume that composites such as all cause mortality, any hospital admission, overall quality of life, or any adverse event, might be more relevant to patients than a particular cause of death or a particular adverse effect. However, such composite outcomes may not provide the information needed for patient care. In practice, each patient has a different risk profile for each of the components of the composite. Even if all cause mortality is the outcome that matters to patients, knowing how the treatment affects specific causes of death is more important for patient management.

It is also tempting to believe that all cause mortality provides the net effect of the treatment[17] and so provides an attractive summary for policy makers. However, conclusions made on the basis of the effect on a composite outcome are applicable to the population where the trial took place and are not readily generalisable to other populations.[10] The net effect on a composite outcome in a given population depends on the relative frequency of the component outcomes. Even if the effects on specific component outcomes are the same across populations, the net effect on the composite will not be the same if the distribution of component outcomes varies. In particular, the net effect of the composite outcome will be diluted in populations where the components that are not causally related to the treatment were common. In a multisite or multinational trial, the net effect might vary by hospital or country even if the specific effects remain constant.

In clinical practice, doctors often prefer highly sensitive diagnostic tests that have a low proportion of false negative results to ensure that few patients are denied an effective treatment. However, in clinical trials, it is important to use highly specific diagnostic tests that have a low proportion of false positive results to prevent dilution of the treatment effect.

Contributors: DP-M and IR did the main writing and LS and TvS made the corrections. All authors critically revised the draft for important intellectual content, and gave final approval of the version to be published. DP-M is the guarantor for the study.

Competing interests: All authors read and understood the BMJ policy on declaration of interests and have no relevant interests to declare.

Provenance and peer review: Not commissioned; externally peer reviewed.

1 ICH Expert Working Group. Statistical principles for clinical trials E9. International conference on harmonisation of technical requirements for registration of pharmaceuticals for human use. 1998. Technical report. www.ich.org/products/guidelines/efficacy/efficacy-single/article/statistical-principles-for-clinical-trials.html. Accessed 26-Sept-2013.
2 Moher D, Hopewell S, Schulz KF, Montori V, Gotzsche PC, Devereaux PJ, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ* 2010;340:c869.
3 Cordoba G, Schwartz L, Woloshin S, Bae H, Gotzsche PC. Definition, reporting, and interpretation of composite outcomes in clinical trials: systematic review. *BMJ* 2010;341:c3920.
4 Ferreira-Gonzalez I, Busse JW, Heels-Ansdell D, Montori VM, Akl EA, Bryant DM, et al. Problems with use of composite end points in cardiovascular trials: systematic review of randomised controlled trials. *BMJ* 2007;334:786.
5 Ferreira-Gonzalez I, Permanyer-Miralda G, Busse JW, Bryant DM, Montori VM, Alonso-Coello P, et al. Composite endpoints in clinical trials: the trees and the forest. *J Clin Epidemiol* 2007;60:660-1.
6 Ferreira-Gonzalez I, Permanyer-Miralda G, Busse JW, Bryant DM, Montori VM, Alonso-Coello P, et al. Methodologic discussions for using and interpreting composite endpoints are limited, but still identify major concerns. *J Clin Epidemiol* 2007;60:651-7, discussion 658-62.
7 Freemantle N, Calvert M, Wood J, Eastaugh J, Griffin C. Composite outcomes in randomized trials greater precision but with greater uncertainty? *JAMA* 2003;289:2554-9.
8 Freemantle N, Calvert M. Weighing the pros and cons for composite outcomes in clinical trials. *J Clin Epidemiol* 2007;60:658-9.
9 Montori VM, Permanyer-Miralda G, Ferreira-Gonzalez I, Busse JW, Pacheco-Huergo V, Bryant D, et al. Validity of composite end points in clinical trials. *BMJ* 2005;330:594-6.
10 European Network for Health Technology Assessment. Endpoints used for REA of pharmaceuticals: composite endpoints. 2013. www.eunethta.eu/outputs/methodological-guideline-rea-pharmaceuticals-composite-endpoints.
11 Rothman KJ, Lash TL, Greenland S. Modern epidemiology. Lippincott Williams and Wilkins, 2012.
12 Kessler KM. Combining composite endpoints: counterintuitive or a mathematical impossibility? *Circulation* 2003;107:e70.
13 Cockburn J, Behan M, de Belder A, Clayton T, Stables R, Oldroyd K, et al. Use of troponin to diagnose periprocedural myocardial infarction: effect on composite endpoints in the British Bifurcation Coronary Study (BBC ONE). *Heart* 2012;98:1431-5.
14 Rodgers A, McMahon S. Systematic underestimation of treatment effects as a result of diagnostic test inaccuracy: implications for the interpretation and design of thromboprophylaxis trials. *Thromb Haemost* 1995;73:167-71.
15 CRASH-2 Collaborators. Effects of tranexamic acid on death, vascular occlusive events, and blood transfusion in trauma patients with significant haemorrhage (CRASH-2): a randomised, placebo-controlled trial. *Lancet* 2010;376:1-10.
16 HPS Collaborative Group. MRC/BHF Heart Protection Study of cholesterol lowering with simvastatin in 20,536 high-risk individuals: a randomised placebo-controlled trial. *Lancet* 2002;360:7-22.
17 Lubsen J, Kirwan B. Combined endpoints: can we use them? *Stat Med* 2002;21:2959-70.

**Accepted:** 25 October 2013

Cite this as: *BMJ* 2013;347:f6782

### Summary points

- The use of non-specific or composite outcomes could increase the proportion of events that are not causally related to the intervention
- Including events that are not causally related to the intervention will dilute the estimated effect towards the null and can result in more precise estimates of the wrong effect and a reduction of the power in the trial
- Effects on composite outcomes may not provide the information needed for individual patient care
- Conclusions made on the basis of the effect on a composite outcome are not readily generalisable to other populations

# Table

Table 1| **Results for an example trial of β blockers in chronic obstructive pulmonary disease (COPD)**

| | β blocker (n=775)* | | Control (n=775)* | | | |
|---|---|---|---|---|---|---|
| **Outcome** | **No of patients/events** | **Risk/rate†** | **No of patients/events** | **Risk/rate†** | **Relative risk/rate† (95% CI)** | **P value** |
| **Hospital admission** | | | | | | |
| Any cause | 272 | 0.35 | 302 | 0.4 | 0.90 (0.79 to 1.02) | <0.12 |
| COPD related | 121 | 0.16 | 151 | 0.2 | 0.80 (0.65 to 0.99) | <0.045 |
| **Adverse events** | | | | | | |
| Any shortness of breath | 39 746 | 0.1442 | 39 444 | 0.1431 | 1.01 (0.99 to 1.02) | <0.29 |
| Severe bronchospasm | 378 | 0.0014 | 76 | 0.0003 | 4.96 (3.91 to 6.40) | <0.0001 |

*275 575 patient days of follow-up.

†Hospital admissions are described as risks and adverse events as rates.

# Figures

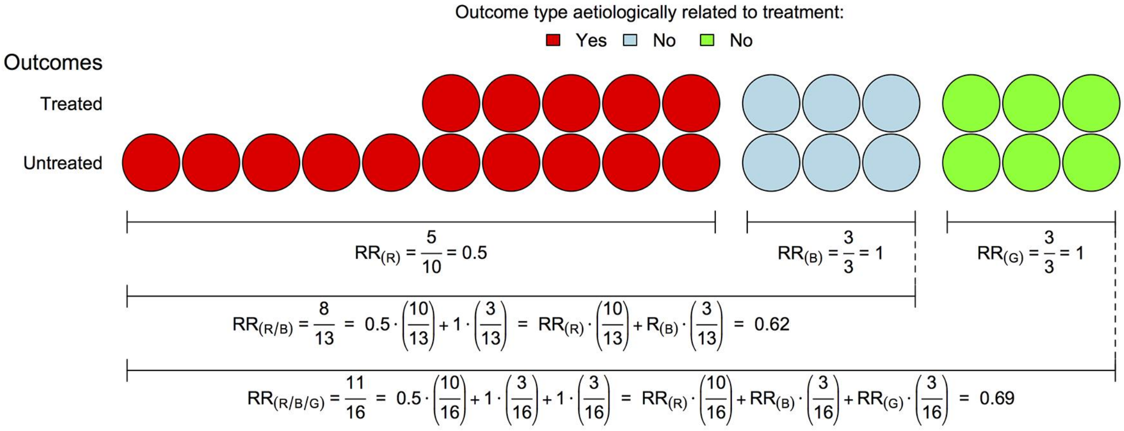
Outcome type aetiologically related to treatment:
🟥 Yes 🟦 No 🟩 No
Outcomes
Treated
Untreated
$RR_{(R)} = \frac{5}{10} = 0.5$
$RR_{(B)} = \frac{3}{3} = 1$
$RR_{(G)} = \frac{3}{3} = 1$
$RR_{(R/B)} = \frac{8}{13} = 0.5 \cdot \left(\frac{10}{13}\right) + 1 \cdot \left(\frac{3}{13}\right) = RR_{(R)} \cdot \left(\frac{10}{13}\right) + R_{(B)} \cdot \left(\frac{3}{13}\right) = 0.62$
$RR_{(R/B/G)} = \frac{11}{16} = 0.5 \cdot \left(\frac{10}{16}\right) + 1 \cdot \left(\frac{3}{16}\right) + 1 \cdot \left(\frac{3}{16}\right) = RR_{(R)} \cdot \left(\frac{10}{16}\right) + RR_{(B)} \cdot \left(\frac{3}{16}\right) + RR_{(G)} \cdot \left(\frac{3}{16}\right) = 0.69$

**Fig 1** Example of composite outcome, only one part of which (red outcome) is related to treatment. We assume the same number of patients in each arm of the trial. (RR= relative risk, R=red, B=blue, and G=green)

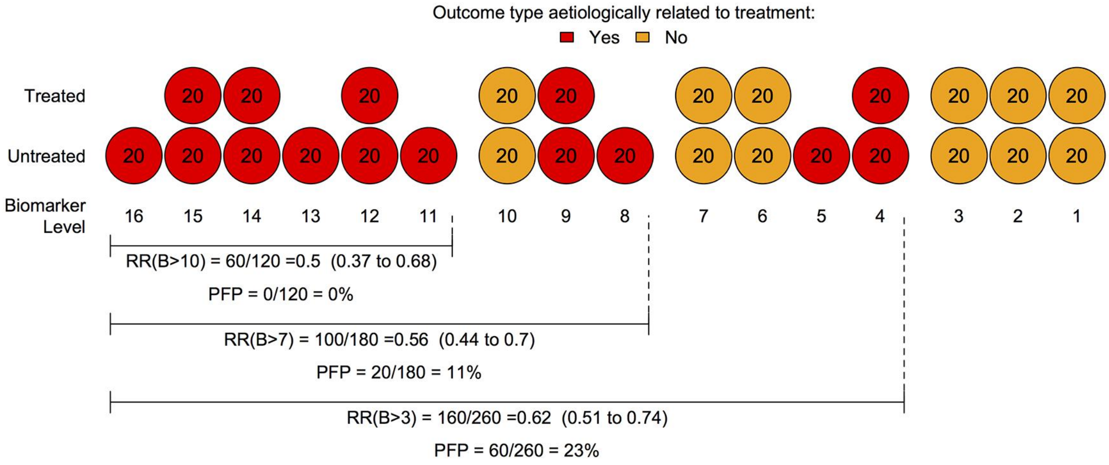
Outcome type aetiologically related to treatment:
🟥 Yes 🟧 No
Treated
Untreated
Biomarker Level
20
16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1
RR(B>10) = 60/120 =0.5 (0.37 to 0.68)
PFP = 0/120 = 0%
RR(B>7) = 100/180 =0.56 (0.44 to 0.7)
PFP = 20/180 = 11%
RR(B>3) = 160/260 =0.62 (0.51 to 0.74)
PFP = 60/260 = 23%

**Fig 2** Example of misclassification of outcomes in a randomised controlled trial based on biomarker level, 20 patients per circle and assuming a total of 2000 patients randomised to each arm. The effect of treatment on aetiologically related outcomes (red circles) is relative risk=0.5. (RR= relative risk, B=biomarker, PFP=proportion of false positive results)

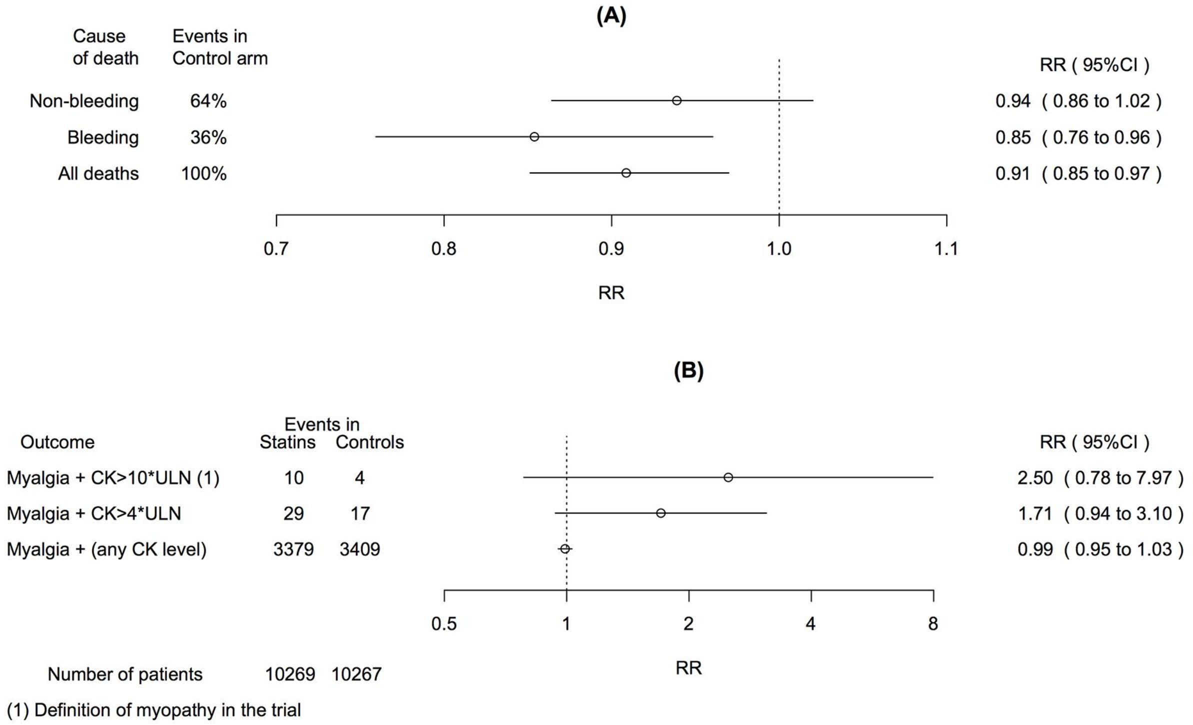

**Fig 3** Examples of composite outcomes diluting effects of treatment from (*a*) the CRASH-2 trial[15] and (*b*) the Heart Protection Study[16] (CK=creatine kinase, ULN=upper limit of normal)