A brief summary of the process on the Cochrane review on vitamin C for atrial fibrillation

Harri Hemilä **2017-11-28**
http://www.mv.helsinki.fi/home/hemila

In 2015, we wrote a protocol with a colleague of mine for vitamin C and AF.
http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD011471/full

We submitted the manuscript for Cochrane review and it was finally rejected with the following arguments on the next page (p. 2).

The final editorial office comments and our responses to them are on the following sheets.
It was not clear to us what was the actual reason for rejection.
We responded to the new comments and our responses are marked by blue.
The greatest issues seemed to be:

1) our subgroup analysis (p. 4-7) and

2) our argument that double-blinding implies that there must be allocation concealment, otherwise there cannot be blinding after randomization (p. 11-12).

The editor did not point out problems in our reasoning on the above issues, nor any specific problems with other issues. Thus, the actual reasons are not described in the rejection letter.

For issue 1) we showed that "Our subgroup analyses fully or partly satisfy 10 out of 11 criteria listed by Sun et al. and the interaction P values are particularly low. Therefore we do not consider that there is basis to consider that our subgroup findings are "extremely unlikely".

For issue 2) the reasoning is very simple:
Our reasoning followed basic logic, called "syllogism".
The standard example of syllogism is as follows:
All men are mortal, Socrates is a man, therefore Socrates is mortal.

Our argument is:
Double-blinding means that over all the time points of a trial, from the very beginning to the very end, the participants and the researchers are unaware about the treatment of a particular patient in the controlled trial.
Allocation is a time point within a controlled trial (at the very beginning).
Therefore, the existence of double-blinding implies that participants and researchers are unaware about the treatment at the time point of allocation.
Thus, double-blinding implies allocation concealment.
But not vice versa: mortality of Socrated does not logically imply that all men are mortal.

The draft that was intended for the Cochrane was published in BMC Cardiovascular Disorders:
https://doi.org/10.1186/s12872-017-0478-5

19 May 2016

Dear Prof Hemilä and co-authors

We discussed your submission of review Vitamin C for treatment and secondary prevention of atrial fibrillation among the editors of the Heart Group and the Cochrane Central Editorial Unit. We are sorry to report that it is not acceptable for publication. The quality expected of Cochrane reviews has not been met.

You had twice the opportunity to revise the review and there are still major concerns (analysis, interpretations of your findings, inconsistencies, inadequate use of GRADE, missing information in Summary of Findings table, risk of bias judgements).

The published protocol will be withdrawn from the Cochrane Library and we reserve the right to offer the review title to another author team. You are entitled to seek publication with another journal. This follows our withdrawal policy as outlined here: http://heart.cochrane.org/withdrawal-policy.

We hope to have your continued support for the work of the Cochrane Heart Group.

Best wishes,

JP Casas
Co-ordinating Editor
Cochrane Heart Group

RESPONSES TO REVIEWER COMMENTS DATED 2016-3-1

Vitamin C for treating atrial fibrillation
New review, 1308
Draft submitted: 20/9/2015
Editorial comments (no peer review, no CE) sent to authors: 24/9/2015
Revisions received: 19/10/2015
Further comments (no peer review) sent to authors: 12/2/2016

New comments [by the Cochrane editors] are marked by yellow

Responses by **HH and TS March 8, 2016**
Responses **HH+TS** are without color after the yellow new comments.

Abstract

1. Main results: Please report on the overall effects on all of your primary outcomes (not subgroup analysis and secondary outcomes). Currently not reported are all cause mortality and adverse effects.
   Authors' response: Done
   Further comment: Here you state that no adverse effects were **observed**. This is contradictory to the statement under effects of intervention where you state that no concerns of adverse effects were **reported**. Please clarify.

HH+TS: we changed the word to "reported".

2. **New comment:** Data collection and analysis: this section of the abstract should give a succinct summary on the rigors of the methods for data collection and analysis used in the review. It should also include GRADE and how it rates the quality of evidence as this is a vital part of data collection and analysis

HH+TS: we added more details

3. **New comment:** Main results: refrain from the using the words 'significantly decreased' and 'the number needed to treat to benefit (NNTB)' as these might mislead the reader and does not convey the true sense of the result findings. In place of these, report the magnitude and precision of the estimated effect around each pre-specified outcome from your summary of findings table and describe the quality of evidence as indicated from GRADE ratings e.g. (RR xxx, CI (xxx to xxx); nos. of studies; high or moderate or low or very low quality of evidence). Report the main outcomes of interest irrespective of the strength of evidence. As a general approach, outcomes important enough to feature in the Summary of Findings table should be considered for the abstract and vice versa. This way of reporting should apply throughout the review.

HH+TS: NNTB and significant removed.

"Report the main outcomes of interest irrespective of the strength of evidence"

Does this mean ICU stay and mechanical ventilation?
We added them.

" describe the quality of evidence as indicated from GRADE ratings e.g. (RR xxx, CI (xxx to xxx); nos. of studies; high or moderate or low or very low quality of evidence)"

Our reporting was already in this format.
We do not understand that comment.

4. **New comment**: Main results: "The absolute duration of hospital stay was shortened by 0.73 days (95% CI 0.45 to 1.00 days; participants = 1399; studies = 9; I2 = 57%; moderate quality of evidence)". It is the 'duration of hospital stay' not 'absolute duration of hospital stay'. The hospital duration was less by 73% not 0.73 days. Please always report magnitude and precision of the estimated effect and describe the quality of evidence as indicated from GRADE ratings e.g. (RR xxx, CI (xxx to xxx), high or moderate or low or very low quality of evidence) OR (MD xxx lower (xxx lower to xxx higher, high or moderate or low or very low quality of evidence) compared to xxx as the case may be.

HH+TS: There is some strange misunderstanding by the above reviewer. Our Analysis 2.2 shows that the effect of vitamin C on hospital stay is "-0.73 days" and surely not -73%.

Since the relative effect is more generalizable, we are primarily using the relative effect. The relative effect of vitamin C on hospital stay is "-11.39%" in Analysis 2.1. "Absolute duration" means that we are counting the days. We rewrote that sentence to state that "In terms of days, ..."

English is a second language for us and we do not know if this is optimal way to write, but we consider that both the "relative effect" and the "absolute effect" (in days) for hospital stay should be mentioned in the Abstract.
Please, give suggestions if you consider that this version is no clear.

"describe the quality of evidence as indicated from GRADE ratings e.g. (RR xxx, CI (xxx to xxx), high or moderate or low or very low quality of evidence)"

Our reporting was already in this format.
We do not understand that comment.

5. **New comment:** Main results: It doesn't seem appropriate to mention post-hoc subgroup findings in the abstract. These subgroup findings meet only very few credibility criteria (see Sun et al. BMJ 2010;340:c117).

The Sun et al. paper:
https://doi.org/10.1136/bmj.c117

HH+TS: We removed the mention of the subgroup findings from the Abstract

However,
it is quite incorrect to claim that our subgroup findings
"meet only very few credibility criteria" of the paper by Sun et al.

Sun et al,. write about the level of statistical significance in interaction P value as follows (p. 852):

> "When examining subgroup hypotheses, one must address the likelihood that the differences in effects

can be explained by chance. The statistical approach that addresses this issue is called a test for interaction (the interaction meaning that the treatment effect differs across subgroup categories). The null hypothesis of the test for interaction is that no difference exists in the underlying true effect between subgroup categories. The lower the P value, the less likely it is that chance explains the apparent subgroup effect. Inevitably, the choice of a threshold for the P value involves subjective judgment. Rather than use of a threshold, a preferable way of assessing the P value is that as it gets smaller, the subgroup hypothesis becomes increasingly credible: we can be sceptical of any hypothesis with a P value of greater than 0.1, begin to consider the hypothesis if the P value is between 0.1 and 0.01, and take the hypothesis seriously when P values reach 0.001 or less"

Thus, the "interaction P-value" is an essential concept in the Sun et al paper in BMJ.

They write (above) that
"The lower the P value, the less likely it is that chance explains the apparent subgroup effect … "
"... a preferable way of assessing the P value is that as it gets smaller, the subgroup hypothesis becomes increasingly credible"

They describe that if the P-value for interaction is over 0.1 it is difficult to become excited.
On the other hand, in their view, researchers should
"take the hypothesis seriously when P values reach 0.001 or less"

In our Analysis 1.5, P = 0.0004 for the subgroup differences over the three subgroups, thus it is smaller than the level Sun et al suggest to "take the hypothesis seriously"

In our Analysis 2.3, P = 0.003 for the subgroup differences over the two subgroups, thus it is smaller than the range 0.1 to 0.01 where Sun et al suggest to "begin to consider the hypothesis".

The importance of the other criteria depends on the test of significance in the subgroup comparison test. Say, if interaction P = 0.01, then we are obviously interested in how many variables had been tested. In RCTs it is common that a few dozen or a hundred variables are collected at baseline and multiple comparison problem in such a context easily explains subgroup P = 0.01. However, subgroup difference at level P<0.001 is not easily explained by a few dozen comparisons, which gives the reason for Sun et al. comment "take the hypothesis seriously when P values reach 0.001 or less"".

1) Thus, both the above subgroup analyses (in particular Analysis 1.5) satisfies the Sun et al (BMJ) criteria for small interaction P-value.

In addition, we can look at the list of other criteria by Sun et al.:

2) Baseline: the subgroups that we analyze are defined at baseline

3) Specified a priori:
The dosage of vitamin C was explicitly specified a priori in our protocol as "dosage of vitamin C as subgroup variables" and the iv versus po kind of difference is therefore explicitly prespecified.

In our manuscript we write: “we planned that if there are suitable data available, we were interested in the potential role vitamin C status as a subgroup variable. None of the included studies reported vitamin C status of the patients. After seeing the included studies we observed that there was substantial divergence in the effects of vitamin C in the US studies and in the Iran studies and we decided to carry out a post hoc subgroup analysis by the countries so that we contrasted US and non-US studies, and Iran and non-Iran studies. This country-based subgroup analysis is not unambiguously inconsistent with the concept of dietary vitamin C intake influencing the effects of vitamin C supplementation, since it is possible that the average intake is lower in Iran than in the USA. However, we did not find data about average dietary vitamin C intake in Iran.”
Thus, although we did not prespecify a comparison of non-US vs US, that comparison is not far from what we did prespecify.

4) Direction: smaller effect in the USA is anticipated by the BMJ paper to which we refer to and greater effect of intravenous vitamin C is expected since intravenous administration causes higher levels of the vitamin in body.

5) Small number of effects tested: we have not been doing extensive testing of numerous variables so that the two subgroups would be picked from a multitude. This is obvious since the number of characteristics reported of trials is quite limited. The situation is different in the case of original RCTs in which dozens or hundreds of baseline variables can be collected and a within trial subgroup can therefore suffer from a serious multiple comparison problem. In such a case the number or subgroup analyses can be very large.

6) Independence: it is highly likely that the comparison of iv against po vitamin C is an independent in the sense Sun et al describe in their paper. Country is an umbrella for numerous characteristics, and subgroup analysis by country is not independent in the same sense as the method of vitamin C administration.

7) Size of the effect: In Analysis 1.5 the smallest estimate is 7% difference between vitamin C and control (USA) and the largest is 63% difference (non-US oral) which is a large difference in the pooled estimates In Analysis 2.3 the difference between 6% and 17.7% reduction in hospital stay is also a large differences between the pooled estimates.

8) Consistency across studies: In Analysis 2.3, all the point estimates for oral studies are larger than all the point estimates for iv studies; in Analysis 1.5 all iv studies have point estimate larger than all iv studies, but 2 US studies have point estimates within the region covered by the non-US oral studies, however, the 2 studies are very small and noninformative.
Thus there is very much consistency in the findings

9) Closely related outcomes: “The consistency of the subgroup effect across outcomes enhances its credibility” (p 853)
Occurrence of POAF may influence the stay at hospital and therefore the effect of vitamin C on both outcomes is consistent with the criteria that closely related outcomes are influenced

10) Indirect evidence: given the same dose of vitamin C, it is both obvious, but also documented experimentally that intravenous vitamin C causes greater levels of vitamin C in plasma. Also, a meta-analysis which we refer to in our manuscript reported that for many outcomes the studies that were carried out in developing countries found greater effects than studies carried out in developed countries.

Thus, it is quite incorrect to state that our subgroup findings
“meet only very few credibility criteria” of the paper by Sun et al.
10 criteria out of 11 is not “only very few”

Only one criteria
"comparisons within studies rather than between studies"
is clearly not satisfied.

According to Sun argumentation in their BMJ paper - as pointed out above - the smaller the P-value -- the more we should take the finding seriously. Thus, the smaller the P-value, the less reasonable it is to explain the P-value by the multiple comparison problem. Our interaction P values are particularly small as mentioned above.

Sun et al do not oppose subgroup analysis. They oppose the Douglas Altman-type of view that all subgroup analyses are evil (and refer to ref 6 in their paper) and Sun et al write "This "yes" versus "no" polarised approach is undesirable and destructive" (p 850).

Furthermore, Sun et al do not object post hoc subgroup analyses.
In contrast, they describe their own post hoc subgroup analysis (p. 853; bold added by us):
"... We subsequently used the trial data to **explore five additional hypotheses**, one of which suggested that reamed nailing was superior in current smokers (RR 0.68, 95% CI 0.50 to 0.92) and unreamed nailing better in others (that is, ex-smokers and lifetime non-smokers) (RR 1.56, 95% CI 1.04 to 2.36, interaction P=0.001, fig 1)."

Thus, in their post hoc subgroup analysis, Sun calculate that the interaction P = 0.001.
In our Analysis 1.5, interaction P = 0.0004 and in our Analysis 2.3, interaction P = 0.003. The former is much smaller than the interaction P that Sun et al. considered worth to report in an educational paper on subgroups in BMJ, and the latter is not much larger than their interaction P.

Sun et al write "Treating the likelihood that a subgroup effect is real as a continuum reflects the nature of the uncertainty … the greater the extent to which criteria are met, the more likely the subgroup effect is real. When summarising the strength of the subgroup inferences, one can imagine—and possibly apply—a visual analogue scale with anchors of "highly plausible" and "extremely unlikely" " (p. 853).

Our subgroup analyses fully or partly satisfy 10 out of 11 criteria listed by Sun et al. and the interaction P values are particularly low. Therefore we do not consider that there is basis to consider that our subgroup findings are "extremely unlikely".

6. **New comment**: Authors' conclusion: your conclusions do not reflect your main result findings. Nowhere in the main results was evidence provided to support your 2g/day recommendation. Also, please refrain from making any recommendation especially when it's not evidence based. Your conclusion must always reflect your main results findings and the quality of evidence (according to GRADE) and how these translate into needing 'further research'.

HH+TS: We removed recommendations

Plain language summary

1. Please use sub-headings (review question, background, study characteristics, key results, quality of the evidence) as recommended here: http://editorial-unit.cochrane.org/sites/editorial-unit.cochrane.org/files/uploads/PLEACS_0.pdf.
   Authors' response: we rewrote the PLEACS according to the instructions

Further comment: Thanks for revising. The review question seems unclear – effect of Vit C on AF in people at high risk of AF. I suggest you rephrase to 'effect of Vit C on AF and for the secondary prevention of AF in people at high risk'.

HH+TS: Done

2. **New comment**: Emphasis should be on the uncertainty in the effect sizes rather than reporting percentages.

HH+TS: We asked for more detailed instructions and March 2 Nicole Martin wrote:

> Without describing the uncertainty surrounding the absolute effect size, the presentation of the results is misleading. In addition to the absolute effects, you could just add a statement to say that the results are subject to some uncertainty surrounding this effect. For example, the way it currently reads is that Vitamin C definitely decreases the risk of AF by 34%. But it could be another percentage.

HH+TS: We do not believe than an average reader of our PLS considers that 34% is intended to be an accurate estimate.
However, to avoid misunderstandings,we removed the percentages from the PLS.
We hope that we have correctly understood your instructions.

Methods

1. Types of studies: A specification 'which **measured** AF as an outcome' has been added (as compared to the protocol). The **reporting** of a particular outcome should not influence the inclusion/exclusion of a study. It looks as though, for example, Dingchao 1994 meets the inclusion criteria and reports two of the outcomes of your interest. However, it's been excluded as it doesn't report on occurrence of AF. Please revisit your excluded studies and include any which fit the inclusion criteria (regardless of **reporting** on the main outcome or not).
   Authors' response: there are two different issues in the text above: "which measured" and "the reporting".
   If a study measures an outcome, such a study should be included irrespective of the reporting of the result. We follow this reasoning so that we describe in our Results section that Samadikhah and Healy measured hospital stay, and the former also ICU stay, but they did not report the data since the effect of vitamin C was not significant.
   The Handbook writes (5.1.2):
   "Outcomes usually are not part of the criteria for including studies: a Cochrane review would typically seek all rigorous studies (e.g. randomized trials) of a particular comparison of interventions in a particular population of participants, irrespective of the outcomes measured or reported.
   However, some reviews do legitimately restrict eligibility to specific outcomes"
   and (5.4.1):
   "In some circumstances, **the measurement of certain outcomes may be a criterion for including studies into a review**, for example when the intervention is aimed at preventing a particular outcome. However, **reporting** of outcomes should rarely determine eligibility of studies for a review."
   HH+TS: Bold is added by us.
   In our review, the intervention is aimed at preventing AF.
   The use of AF in the inclusion criteria is consistent with the above Handbook comment.
   We could have planned in our protocol that our focus is on cardiac surgery. In such a case we would have planned our searches to find studies with cardiac surgery.

However, when we have planned our protocol to focus on people with high risk of AF, the condition AF should be in the search methods. We are including studies other than cardiac surgery since high risk of AF is not restricted to cardiac surgery. Heavy exercise increases risk of AF, high blood pressure increases risk of AF, one of the included studies used participants after cardioversion, and such patients have a high risk of AF.

We added 'which measured AF as an outcome' to our review when we realized that our search requires AF, but our "types of studies" in the Protocol did not.
Thus, we added that term to make the search consistent with the types of studies.
Term "high risk of AF" has a clear meaning but it is difficult to make it operational without introducing AF as a search term.
We are ready to revise our "types of studies" description if you have further suggestions.
However, it is not simple to define "high risk of AF" without mentioning the term AF, but still covering such conditions. It does not make any sense to search all studies on exercise and read if the study happened to report on AF. If a meaningful measurement of AF has been done in a study, it is essentially always listed as a keyword.
Now we keep the AF.
This is consistent with the Handbook (5.4.1):
"In some circumstances, the **measurement** of certain outcomes may be a criterion for including studies into a review, for example when the **intervention is aimed at preventing a particular outcome**."
There is no description in the Dinchao report whether the study was randomised or not and therefore we cannot include it even though we would be satisfied with the secondary outcomes.
We incorrectly stated that the study was RCT, we corrected that (RCT is a loose word often used to mean trials that may also be quasi-randomised such as alternative allocation). Nevertheless, the hospital stay and ICU stay are reasonable issues in our Discussion section.
Further comments: Thanks for your explanations and revising the reason for exclusion for Dinchao. My concern about adding AF (outcome) as an inclusion criteria is that you may exclude studies which don't report on this outcome but other outcomes you are interested in. However, if you decide to continue with this approach, please be more specific, ie RCTs… which measured incidence of AF in secondary prevention trials (your first primary outcome). That will exclude all your treatment trials.

HH+TS: We asked for more detailed instructions and March 2 Nicole Martin wrote:

You didn't pre-specify in the protocol that one of the inclusion criteria is to measure AF as an outcome. As this is now in the review, this solves the problem. Please make this change in inclusion criteria clear under 'differences between protocol and review' and justify why this was done.

HH+TS: We added a description of the changes to the "differences" section.

2.

Types of studies/types of outcome measures: including secondary prevention trials seems contradictory to the title of the review which seems to focus on treatment. Please explain.
Authors' response: could you please give closer instructions.

HH+TS: Old text removed from here to shorten this text

3. **New comment**: The outcomes reported are different from outcomes pre-specified which makes the report difficult to understand.

HH+TS: We asked for more detailed instructions and March 2 Nicole Martin wrote:

> Methods: incidence of AF vs Results: Occurrence of POAF and recurrence of AF after cardioversion
> Methods: Secondary prevention trials with POAF patients: length of hospital stay vs Results: Length of hospital stay of patients undergoing cardiac surgery
>
> Methods: Secondary prevention trials with POAF patients: length of ICU stay vs Results: Length of ICU stay of patients undergoing cardiac surgery
>
> Methods: Secondary prevention trials with POAF patients: length of mechanical ventilation vs Results: Length of mechanical ventilation in patients with cardiac surgery
>
> Methods: all-cause mortality vs Results: mortality
>
>
> They should also be reported in the same order as planned, eg all-cause mortality before length of hospital stay, throughout the review.

HH+TS: We modified the titles and reordered the presentation, we hope the revised version is satisfactory

Risk of bias in included studies

1. Allocation bias: You seem to confuse allocation concealment with blinding. Please see Handbook, chapter 8.10. Please check and revise your judgements and explanatory notes accordingly.
   Authors' response: No, we are not confusing allocation concealment and blinding.
   The definition: allocation concealment (Handbook 8.10.2):
   "Allocation concealment … protecting the allocation sequence *before and until* assignment"
   Thus, this time frame for keeping all participants and researchers blinded is from "before" to "until assignment has been done".
   The definition for double (or triple) blinding is that no researchers or participants knows to which group the person is allocated with a time frame from "before allocation" to the "until all the data has been collected and the study terminates."
   Thus, the phenomenon is the same (participants and researchers are kept ignorant of the study groups).
   The time frame of allocation concealment is very short – it is a short period within the time frame of the "double(triple)-blinding."
   The rationale for allocation concealment is that there are treatments that cannot be blinded at the intervention stage, eg in surgery. However, participants in such studies can be allocated in a blinded fashion (patients and researchers are blinded about the allocation). Thus, allocation concealment is a phenomenon that can be done more easily and we can require it eg for studies on surgery.
   Double(triple)-blinding cannot be done for all treatments, but when it can be done, it always leads

to allocation having been concealed (blinded). Otherwise there could not be blinding at the later stages of the trial.
The section to which you referred, writes:
“Thus, allocation concealment up to the point of assignment of the intervention and blinding after that point address different sources of bias and differ in their feasibility.”
That is incorrect or at least misleading.
”Blinding” in double-blind study does not start from ”after that point”. Blinding in double-blinded studies starts from the very beginning, ie. before allocation. Thus, it covers the stage of allocation.
The text also states what we write above:
“Allocation concealment ... can always be successfully implemented regardless of the study topic”
“blinding ... cannot always be implemented”
For the above reasons it is important that we have the two separate concepts.
However, double-blinding implies blinding also at the stage of allocation, when double-blinding has been used.
Further comment: Allocation concealment (selection bias) is different to blinding (performance and/or detection bias). Allocation concealment relates to the issue of being able to guess/or infer the next allocation (treatment/control) and therefore the bias that the next patient is not allocated truly “at random” (compromising the prognostic balance of the comparison groups at the outset through intended or unintended selection). Blinding of patients/care-givers/outcome assessors is a separate issue. Trial authors often use the term double-blinding when e.g. a placebo is used as control intervention; this can be ambiguous in terms of who was actually blinded (see Deveraux et al. JAMA 2001;285(15):2000-3) but it does NOT concern the allocation process.”

HH+TS: Again, that is a misunderstanding from your part.

When a study is double-blinded that “relates to the issue of being able to guess/or infer the next allocation (treatment/control) and therefore the bias that the next patient is not allocated truly “at random”.

If a study is double-blinded, then neither the researcher nor the patient knows in which group that patient was allocated. Neither can either of them know to which group the following patient will be allocated.

The Deveraux paper to which you refer does not deal with this question.
That is a gallup about how do different people interpret the term “double-blinding”.

There are several people involved in RCTs and “double” is not sufficient to describe a complex study. It is preferable to describe specifically that outcome assessment was blinded (if it was), nurses were blinded (if they were), various physicians participating in the study were blinded (if they were) etc.

For example, double-blinding does not necessarily imply that outcome assessment was blinded, therefore the term double-blind -with its history starting some half a century ago- is not useful any more.

Nevertheless, given the standard interpretation that double blinding indicates that patients and the physicians do not know to which group the patient belong at the beginning of the trial -which is the standard answer also in the Deveraux paper (please see their Table) - it is impossible that those people would be aware of allocation, otherwise they would not be blinded.

Allocation concealment is possible in studies that cannot be blinded at the stage of treatment, eg in surgery. However, that does not go other way. If there is double-blinding, it is impossible to have allocation open to any party.

If allocation is open, the people cannot be blinded thereafter, we cannot clean the memory of researchers and patients if they are told about the allocation. Therefore double-blinding indicates that allocation to groups had to be blinded.

If you disagree with this argument, please describe any kind of imaginary study which would be double-blinded but would not have allocation concealment.

The main point of the Devereaux paper is "Our results suggest that authors and journal editors should abandon the terms single, double, and triple blind, and substitute descriptions stating which of the relevant groups were unaware of allocation" (p 2002-2003).
That is not a relevant issue when we consider whether double blinding logically implies (or does not imply) the necessity of allocation concealment.

Effects of interventions

1. P-values can be misleading. Please don't report them. Please also don't use the phrase statistical significance but instead comment on the magnitude of the effect and quality of the evidence.
   Authors' response: We do not agree with you.
   The smaller the P-value, the less likely it is to arise by chance.
   In observational studies small P-values can be irrelevant, since there may be un-identified systematic biases. In such context a reader should be cautious about small P-values.
   However, in RCTs there are no systematic differences between the study groups and therefore the smaller the P-value, the less likely the difference is explained by chance alone. Interpretation of P-values in RCTs is much more clear than in observational studies.
   We assume that the average reader understands that a confidence interval ending close to the null effect is marginally significant or nonsignificant, but most readers do not have any understanding how unlikely are narrow 95%CIs that are far from the null effect.
   Nevertheless, we followed your instruction and removed the P-values from the estimates of effects.
   ok
   Significant:
   In our text, word "significant" is used in different contexts, and we would like more specific suggestions.
   For example, in section "Prevention of AF in high risk patients", we write:
   "Four studies did not find a statistically significant effect of vitamin C."
   If a study does not find a significant effect, one possibility is that there does not exist any effect, but another possibility is that the statistical power of the study was not sufficient to show an effect.
   Therefore the above sentence is relevant, when it is followed by "Three of them were particularly small trials with ≤40 patients" and two of those had just 22 and 24 participants.
   Although the rationale of meta-analysis is to pool the studies and look at the pooled effect, it is also important to look at the individual studies.
   We also use significantly in the following context:
   "The first study reporting benefit of vitamin C against POAF was by Carnes 2001. In that trial, which used historical controls, the incidence of POAF was significantly lower in CABG patients administered vitamin C."
   Although we could show the estimate and its 95%CI, in our view that kind of data is not useful in our review focusing on RCTs. The Carnes study served as the motivation for the later RCTs that we analyze, but its quantitative results are not relevant in our review.
   We also write:
   "In six studies with POAF that found significant benefit of vitamin C, the Number Needed to Treat to Benefit (NNTB) ranges from 4.2 to 6.6"
   It does not make sense to calculate NNT values for nonsignificant differences, and therefore we need to specify our restriction in the sentence above.

In addition, in the above context the text would be confusing if the estimates and CIs were listed.
We also write:
"Samadikhah 2014 and Healy 2010 write in their text sections that the effect of vitamin C on the length of hospital stay and ICU stay was not significant, but they did not report the data."
Thus, Samadikhah and Healy used the "significance" dichotomization and they had decided that they do not report the numerical data. We cannot show or comment on the magnitude of their differences.
In the test of heterogeneity, I-square is currently a useful measure, but that has not made the Chi-square test outdated. We are interested in the probability that the heterogeneity is explained by chance, in addition to the level of heterogeneity.
Furthermore, there are simple algorithms how to calculate the 95%CI for I-square, but RevMan has not implemented them. Thus, there can be a rather large point estimate of I-square, yet the 95% CI of the I-square may extend to null heterogeneity. Thus, this is one context in which we need term "significance" of heterogeneity.
This is discussed in the Handbook (9.6.6) as follows:
"Is there a statistically significant difference between subgroups?
To establish whether there is a different effect of an intervention in different situations, the magnitudes of effects in different subgroups should be compared directly with each other."
Subgroups can be compared by ratio of RRs (ie RRR) and their CIs, but that would be confusing for average readers. Simply testing the significance is easier for the average reader.
In Discussion we write
"We do not consider it likely that the significant effects of vitamin C on the occurrence of AF or on the length of hospital stay might be explained by publication bias."
The word "significant" here means that the differences are not easily explained by chance. In this context of considering biases, the point estimates of our findings are not relevant, but it is useful to emphasize that the calculated differences are not explained by chance when we consider various biases.
The goal of our Results section is to calculate point estimates and 95%CIs for the effects, but in many contexts of Discussion the point estimates are not relevant, and the fact that differences are not easily explained by chance is sufficient. That makes writing and reading easier compared with repeating the estimates and the 95%CIs.
We are well aware of various problems in statistical reporting. In the old days, people dichotomized results to significant and nonsignificant (which was largely explained by the lack of computers), and even nowadays in many instances people overinterpret small P-values, in particular in observational studies in which there is (essentially) always the possibility of residual bias.
Because of the low information content of reporting just the P-value, in 1980s several medical statisticians started to encourage the use of CIs. In the extreme, some of them suggested that P-values should be banned.
However, that kind of conclusion is an exaggeration.
For example Fleiss wrote a short comment to AJPH in which he gave examples when P-values are useful:
http://ajph.aphapublications.org/doi/pdf/10.2105/AJPH.76.5.587
https://en.wikipedia.org/wiki/Joseph_L._Fleiss
"Significant difference" is a practical short phrase in many contexts.
We have deleted a number of "significant" words, but there are many contexts in which it is informative and replacing it with estimates of effect and 95%CI would make the text much less readable.
If you have specific suggestions where we should rewrite, we are ready to consider your suggestions.
Further comment: Instead of discussing significant effects in individual studies, you should focus on the magnitude of the summary effect estimate for an outcome and the quality of the evidence for that outcome as you suggested. Ideally you calculate NNTs from the summary relative treatment effect applied to one or more baseline risks derived from representative patient cohorts (in case

there is no cohort data available you could also use a weighted mean event rate from control groups of included trials).

HH+TS: For essentially all presentations of results in the Results section, we had the GRADE statement, but there were few that did not have. We added the GRADE to those few.

We removed “significant”. However, the consistency of studies is a relevant issue in guiding the considerations. We write that some studies were consistent with the null effect. When some studies find benefit and some do not, it is relevant to consider possible explanations for divergent point estimates. A small study has a wide confidence interval which means that they can be consistent with null effect and large effects, and thus they are simply non-informative. A large study that is negative is thus much more critical compared with a small study that is negative.
We hope that the new version is satisfactory to you.

Summary of findings (SoF) table

1. **New comment**: Please use GRADE ratings and Summary of Findings (SoF) tables to inform the review abstract, Plain Language Summary (PLS), Effects of interventions, Discussion (especially quality of evidence) and your conclusions.
2. **New comment**: The SoF does not meet the standards required of cochrane. Please see below for help:

HH+TS: We modified the SoF table and hope that the revised version is satisfactory.

Our Abstract and Effects of interventions did have the GRADE ratings and we do not understand what the above comment means.

We did have the GRADE evaluation in PLS and we do not understand what the above comment means. We clarified the presentation of GRADE in PLS so that we added the GRADE to the individual outcomes. We hope this revision is satisfactory. If that is not what you mean, could you please describe more detailed instructions.

We added a mention of GRADE to the Discussion section “quality of evidence”

Self management for patients with chronic obstructive pulmonary disease

Patient or population: patients with chronic obstructive pulmonary disease
Settings: primary care, community, outpatient
Intervention: self management[1]
Comparison: usual care

Outcomes	Illustrative comparative risks* (95% CI) Assumed risk usual care	Corresponding risk self management	Relative effect (95% CI)	No of Participants (studies)	Quality of the evidence (GRADE)	Comments
Quality of Life St George's Respiratory Questionnaire. Scale from: 0 to 100. (follow-up: 3 to 12 months)	The mean quality of life ranged across control groups from **38 to 60 points**	The mean quality of Life in the intervention groups was **2.58 lower** (5.14 to 0.02 lower)		696 (7)	⊕⊕⊕⊖ **moderate**[2]	Lower score indicates better quality of life. A change of less than 4 points is not shown to be important to patients.
Dyspnoea Borg Scale. Scale from: 0 to 10. (follow-up: 3 to 6 months)	The mean dyspnoea ranged across control groups from **1.2 to 4.1 points**	The mean dyspnoea in the intervention groups was **0.53 lower** (0.96 to 0.1 lower)		144 (2)	⊕⊕⊖⊖ **low**[3,4]	Lower score indicates improvement
Number and severity of exacerbations[5]	See comment	See comment	Not estimable[6]	591 (3)	See comment	Effect is uncertain
Respiratory-related hospital admissions (follow-up: 3 to 12 months)	**Low risk population**[8]		**OR 0.64** (0.47 to 0.89)	966 (8)	⊕⊕⊕⊖ **moderate**[7]	
	10 per 100	**7 per 100** (5 to 9)				
	High risk population[8]					
	50 per 100	**39 per 100** (32 to 47)				
Emergency department visits for lung diseases (follow-up: 6 to 12 months)	The mean emergency department visits for lung diseases ranged across control groups from **0.2 to 0.7 visits per person per year**	The mean emergency department visits for lung diseases in the intervention groups was **0.1 higher** (0.2 lower to 0.3 higher)		328 (4)	⊕⊕⊕⊖ **moderate**[4]	
Doctor and nurse visits (follow-up: 6 to 12 months)	The mean doctor and nurse visits ranged across control groups from **1 to 5 vists per person per year**	The mean doctor and nurse visits in the intervention groups was **0.02 higher** (1 lower to 1 higher)		629 (8)	⊕⊕⊕⊖ **moderate**[8]	

*The basis for the **assumed risk** (e.g. the median control group risk across studies) is provided in footnotes. The **corresponding risk** (and its 95% confidence interval) is based on the assumed risk in the comparison group and the **relative effect** of the intervention (and its 95% CI).

CI: Confidence interval; OR: Odds ratio;

Also refer to these links on how to report and interpret summary of findings table: http://handbook.cochrane.org/chapter_11/11_presenting_results_and_summary_of_findings_tables.htm ; http://www.cochranelibrary.com/about/explanations-for-cochrane-summary-of-findings-sof-tables.html
Please give clear explanation for downgrading decisions, with a reference to the consideration (e.g. risk of bias or imprecision) and the number of levels downgraded. Please refer to the Cochrane handbook here: http://handbook.cochrane.org/chapter_12/12_2_assessing_the_quality_of_a_body_of_evidence.htm

Please report important patient-related outcomes that were pre-specified in the SoF tables. Outcomes featured in the SoF tables should also be reported in the abstract, PLS, Effects of interventions, Discussion and conclusions.

HH+TS: We modified the SoF table and hope that the revised version is satisfactory.