PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Vitamin D3 supplementation in patients with frequent respiratory tract infections - a randomised, double-blind intervention study
AUTHORS	Bergman, Peter; Norlin, Anna-Carin; Hansen, Susanne; Rekha, Rokeya; Agerberth, Birgitta; Bjorkhem-Bergman, Linda; Ekstrom, Lena; Lindh, Jonatan; Andersson, Jan

VERSION 1 - REVIEW

REVIEWER	Graham Bothamley Honorary Professor and Consultant Physician Homerton University Hospital UK No competing interests
REVIEW RETURNED	01-Aug-2012

THE STUDY	HIV needs to be stated as an exclusion or inclusion criteria. The patient group is pragmatic (attending an immunodeficiency unit), but there are considerable differences between bronchiectasis and the others in terms of the likely endpoints. The dichotomous scoring for ear and sinus symptoms does not reflect the diary.
REPORTING & ETHICS	Randomized and double blind should be included in the title. Baseline data need to be more detailed (see comments to authors) to support generalisability.
GENERAL COMMENTS	This paper describes the effect of 4,000 IU vitamin D3 in patients from a cohort of 84 with antibody deficiency and 56 with frequent respiratory infections (>42 days per year). This is a well-designed trial with clear endpoints. The title should include randomized and double-blind as per CONSORT requirements. Table 1 requires more detail, which could be included as a supplemental table of only those who completed the study. This should identify which patients received supplemental immunoglobulin routinely. Most IgG subclass deficiency does not require supplementation as the deficiency is complemented by the redundancy in IgG subclasses for protein and carbohydrate antigen recognition. Patients with bronchiectasis most commonly provide sputum from which bacterial cultures can be obtained, whereas exacerbations of COPD rarely do so. This supplemental table should therefore indicate from each disease group the number of cultures sent and the number of positive cultures obtained, particularly noting those which were significant (Staph. aureus and fungal cultures).

	This would help exclude the possibility that the greatest effect of vitamin D supplementation was in patients with bronchiectasis, who were the most likely group to provide positive sputum cultures and to receive antibiotics. In order to substantiate the lack of effect of smoking, this should also be included in the enlarged table. Did any patients have concurrent HIV infection? Was this an exclusion criterion? Figure S1 indicates the diary used. However, the ear and sinus score are termed “dichotomous”, but this is not explained in the methods or legend to Table 2, when the diary implies that they would have had a score for each day of >1 symptoms. The supplement explains that the limited number of events meant that logistic regression was used - which seems a better explanation which should be added to the legend. The diary also records number of days of sick leave. This would be a useful variable in exploring the value of vitamin D supplementation. For how long would the additional 3 points be given to a radiological diagnosis of pneumonia? If this were just on the day of diagnosis this would not bias the results significantly, but if for each day of symptoms the score would have been dramatically affected. I would expect that each patient with pneumonia would have received a course of antibiotics so it would again be worth noting the number of courses of antibiotics prescribed for other reasons. Minor points Page 7 line 43. “Thus, a similar but not identical diary was used in the current study….”

REVIEWER	Adit Ginde, MD, MPH Associate Professor of Emergency Medicine University of Colorado School of Medicine Aurora, CO USA
REVIEW RETURNED	09-Aug-2012

THE STUDY	--The group of "increased susceptibility to RTIs" should be briefly defined in the abstract --The primary outcome was the composite score. It is misleading to overemphasize one statistically significant finding of the components (use of antibiotics). Specifically, would recommend removing from the title, as this was not the primary outcome of the trial. It should also be interpreted with appropriate caution as a secondary outcome throughout the rest of the discussion and conclusions. --What is the validity of the composite outcome? Has it been used/validated before? What is the clinical signficance of the differences observed in this population? --For both the inclusion criteria and composite outcome, how are allergic symptoms (which can include many symptoms from the respiratory tract), distinguished from infection. How about chronic symtpoms that are also not infectious (e.g, reflux disease, chronic

	asthma, allergic rhinitis, etc). --This is a select group of patients and the findings very well may not generalize. This should be emphasized in the conclusions. --Unclear when the 25OHD levels were performed (i.e., if during the trial, how were those that assessed outcomes blinded from the results)? --I would like to see more information about the analysis of the primary outcome in the main manuscript. --Unclear whether intention to treat analysis was performed and for which outcomes, given that there were post-randomization exclusions (were those subjects actually included in the analysis)? If all n=70 per group were included, how were missing data handled? --Since AMPs and bacterial cultures were performed at set intervals (unknown whether sick or well), unclear on the value/interpretation of these results are without a consistent protocol for measurement. Also, unclear why the placebo group had nearly double the number of cultures. What was the standardization of measurement for these items? --The CONSORT checklist has an item 95%CIs be provided for each analysis. This is not true for most of the data presented. --The CONSORT checklist has an item that the number of participants (denominator) be included for each analysis-- this has not been done and it is unclear what the denominator is for the analyses (if true intention to treat, should be n=70 per group for each analysis). --In Tables where p-values were provided, would provide the actual value, instead of "NS". --The first supplementatal figure is missing the "n" for the "Completed Vitamin" item.

REVIEWER	Allan Clark Senior Lecturer in Medical Statistics University of East Anglia No competing interests
REVIEW RETURNED	28-Aug-2012

THE STUDY	The outcome measure is a composite measure and it is not stated if this has been validated or any training/education that the patients need to complete this questionnaire. How well are the patients able to identify symptoms related to ongoing respiratory tract infections? The sample size is based on using a p-value of 0.02, this is unusual and needs justification. They should also use this in the paper by reporting 98% confidence intervals and use this significance level for testing. If they do then there discussion and results of the primary outcome are incorrect since it does not meet the 0.02 cut-point. The sample size calculation does not specify the standard deviation used nor does it justify the effect size they are looking for or the

	drop-out rate. In the statisticial analysis they adjust for age, gender, smoking status, type of immune deficiency and co-morbidities. They do not justify these choices or state if they are based on previous research suggesting a link with the outcome measure or imbalance at baseline (which would be a poor justification).
RESULTS & CONCLUSIONS	The results are credible and do relate to the research question. However, they do not appear to be consistent with the sample size calculation. Firstly, they use a different level of significance (5% in the analysis, 2% in the sample size calculation). Secondly, they appear to consider a 23% reduction important in the analysis whereas the sample size calculation implies that the smallest clinically important difference is 30%. Additionally, the unadjusted analysis is based on a Mann-Whitney test whereas the adjusted analysis is based on a regression model with a log-transformed outcome....this makes is impossible to compare the adjusted effect and the unadjusted effect to see if the adjustment made any difference to the effect size. The analysis is also difficult to follow as it is unclear what population the analysis refers to. Normally, it would be the intention-to-treat analysis which would be the main analysis. It is also unclear to me why the authors have chosen to use an ad-hoc imputation method rather than the standard imputation approaches. The statistical properties of this imputation process are not known (for example how does it allow for uncertainity in the impuation process) and hence it shouldn't be used. The statement that it is conservative seems to be made without justification and cannot be true in general, for example if those who dropped out of the placebo group were better than those who remained in the trial then replacing them by the mean of those who remained in the trial would make the placebo group appear worse than they were. The OR for antibiotic use was 0.365 which is losely interpreted to be a "63.5% reduction" it should be clearer that this is a reduction in the odds of use not a 63.5% reduction in use of antibiotics. The subgroup analysis is a little unclear if this is done on the log-transformed outcome or the untransformed. It is clear from the supplementary material what was done, but it is unclear in the main paper.
GENERAL COMMENTS	Overall, I think that this is an interesting study which was well conducted but I think that the analysis of the data can be made much clearer. Some minor points: i) you use "," rather than "." in many places. ii) p-values should not be reported in Table 1. iii) actual p-values should be given rather than "n.s."

VERSION 1 – AUTHOR RESPONSE

Reviewer 1:

--HIV needs to be stated as an exclusion or inclusion criteria.

All patients were screened and found to be negative for HIV. Although not explicitly stated in the first version of the manuscript, being positive for HIV was an exclusion criterion. This information is now added in the manuscript (page 8, line 2).

--The patient group is pragmatic (attending an immunodeficiency unit), but there are considerable differences between bronchiectasis and the others in terms of the likely endpoints.

Some of the patients had a concomitant lung disorder (COPD, bronchiectasis, asthma), which could have had an impact on the observed effect of vitamin D supplementation. To correct for such effects, a multivariate analysis was performed and lung disease was not found to significantly affect the primary endpoint. However, an analysis of lung disorders and bacterial cultures revealed that especially asthmatic patients seemed to have a greater benefit from vitamin D supplementation. However, these data should be interpreted cautiously due to the low number of patients and weak statistical significance (discussed in detail below.). We have added new information in the manuscript on these issues (table S5, Figure S3 and text in both result and discussion sections).

--The dichotomous scoring for ear and sinus symptoms does not reflect the diary.

The original analysis plan focused on the aggregated total score, and did not involve the precise mode of analysing individual score items. For symmetry reasons, we were inclined to analyse these in the same way as the total score, i.e. by adding scores from all days in each period. However, after finishing the data collection we found that ear and sinus symptoms were so scarce (observed in only a minority of the patients) that it proved impossible to achieve a normal distribution necessary for the parametric multivariable regression. In order not to violate the assumptions of the analysis method, we then chose to treat these scores as binary outcomes (present or absent) enabling us to use multiple logistic regression for the confounder adjustment. Although this was not our first choice, one should bear in mind that both the binary coding and the summing of days involves a synthesis of the raw data which consists only of present/absent values for each day in the study.

--Randomized and double blind should be included in the title.

Please, see the new title as stated above.

--Table 1 requires more detail, which could be included as a supplemental table of only those who completed the study. This should identify which patients received supplemental immunoglobulin routinely. Most IgG subclass deficiency does not require supplementation as the deficiency is complemented by the redundancy in IgG subclasses for protein and carbohydrate antigen recognition.

This information has now been collected and is presented in a new table (supplementary table S1).

--Patients with bronchiectasis most commonly provide sputum from which bacterial cultures can be obtained, whereas exacerbations of COPD rarely do so. This supplemental table should therefore indicate from each disease group the number of cultures sent and the number of positive cultures obtained, particularly noting those which were significant (Staph. aureus and fungal cultures). This would help exclude the possibility that the greatest effect of vitamin D supplementation was in patients with bronchiectasis, who were the most likely group to provide positive sputum cultures and to receive antibiotics. In order to substantiate the lack of effect of smoking, this should also be included in the

enlarged table.

We have compiled two new tables (table S4 and table S5) and one new figure (figure S5) on the relation between immunological diagnosis, concomitant lung disease and the number and content of bacterial cultures. In general, the numbers in each subgroup are too small to draw any firm conclusions. However, for patients with subclass deficiency vitamin D supplementation significantly reduced the probability of leaving a bacterial culture during the study period (placebo, 22/24 patients with >1 culture taken versus vitamin D, 12/22 patients, p=0.0065, table S4).
The role of lung disease was also analysed with regards to the number of taken bacterial cultures and positive bacterial cultures (Supplementary figure S3) as well as the microbiology in the positive cultures (supplementary table S5). Overall, the numbers were too small to draw any firm conclusions. However, there was a trend that patients in the vitamin D group with asthma left fewer bacterial cultures and fewer positive cultures compared to asthmatic patients in the placebo group (p=0.080 and p=0.052, respectively, Figure S3). Moreover, vitamin D treated asthmatic patients had significantly fewer fungal findings compared to placebo treated asthma patients (p=0.0476, table S5). For patients with bronchiectasis and COPD there were no significant differences.

Although the effect is small and the number of patients is limited this new information could have clinical implications regarding which patients that could have the most benefit of vitamin D supplementation. We have included this information in the main text (page 15, page 18). Finally, baseline data on smokers is included in table 1.

--Did any patients have concurrent HIV infection? Was this an exclusion criterion?

No patient had HIV and this would have been an exclusion criterion, although not explicitly stated (now mentioned on page 8, line 2).

--Figure S1 indicates the diary used. However, the ear and sinus score are termed “dichotomous”, but this is not explained in the methods or legend to Table 2, when the diary implies that they would have had a score for each day of >1 symptoms. The supplement explains that the limited number of events meant that logistic regression was used - which seems a better explanation which should be added to the legend.

--We agree that this approach is better and the legend has now been rewritten as suggested (legend, table 2).

--The diary also records number of days of sick leave. This would be a useful variable in exploring the value of vitamin D supplementation.

The parameter: ”numbers of days off work or school” was included to study the possible economic impact of vitamin D supplementation. However, it is important to note that this parameter was never included in the predefined analysis plan for the primary endpoint. We have now analysed these data and there is no effect of vitamin D supplementation on the number of days off work or school. However, it is important to consider that there are a number of problems with this analysis. First, almost 40% of the patients in our study do not work or go to school, mainly because of a permanent or long term sick leave. Thus, a zero score could mean either no absence or that this variable did not apply to the individual patient. Reciprocally, many of our patients have symptoms and still go to work, which is a result of personal attitude towards work as well as a consequence of the Swedish Social Security System with a generally strict view upon sick leave. Thus, there is a severe risk that all these factors would severely flaw the analysis of sick leave as a read-out for vitamin D mediated effects. To conclude, since we strongly believe that this parameter does not reflect the true infection status of the patient we have chosen not to include this information in the main manuscript.

--For how long would the additional 3 points be given to a radiological diagnosis of pneumonia? If this were just on the day of diagnosis this would not bias the results significantly, but if for each day of symptoms the score would have been dramatically affected.

Pneumonia was given 3 points extra for a standard-period of 7 days, i.e. 21 extra points per pneumonia. However, the number of X-ray verified pneumonias was quite small and evenly distributed between the two study groups (5 in the vitamin D group and 6 in the placebo group). Thus, we believe that the fixed algorithm of assigning a pneumonia 21 extra points did not greatly affect the main outcome.

--I would expect that each patient with pneumonia would have received a course of antibiotics so it would again be worth noting the number of courses of antibiotics prescribed for other reasons.

The use of antibiotics was analysed using a dichotomous variable for the study period (due to non-normally distributed data). A detailed analysis of all patients with verified pneumonia revealed that all these patients (n=11) did use antibiotics also for other respiratory tract infections during the study period. Thus, excluding all pneumonias would not result in any shift of dichotomous variable regarding antibiotic use for any of these patients.

Minor points

--Page 7 line 43. “Thus, a similar but not identical diary was used in the current study….”

We mean that patients are trained and used to assess their infectious status since a simplified version of the study diary has been used at the clinic for more than 20 years. The first published example was already in 1991 when Ann Gardulf and coworkers used a diary for infectious symptoms and adverse events connected to subcutaneous treatment with immunoglobulins (Gardulf et al, Lancet, 1991). More information has been added (page 7, second paragraph).

Reviewer 2:

--The group of "increased susceptibility to RTIs" should be briefly defined in the abstract.

We have added information for this group in the abstract (>4 bacterial RTIs/year). (Abstract, under “participants”).

--The primary outcome was the composite score. It is misleading to overemphasize one statistically significant finding of the components (use of antibiotics). Specifically, I would recommend removing it from the title, as this was not the primary outcome of the trial. It should also be interpreted with appropriate caution as a secondary outcome throughout the rest of the discussion and conclusions.

We agree and have changed the title and abstract according to the suggestion of the reviewer. In addition, we have added information in the discussion-part where we specifically discuss that the antibiotic parameter has a major impact on the primary endpoint. Thus, we have followed the reviewer’s advice and have written the interpretation of the antibiotic parameter in a more cautious manner. (new title, abstract, page 15, final paragraph).

--What is the validity of the composite outcome? Has it been used/validated before?

A similar diary has been used to monitor these types of patients for more than 20 years. The first use of a diary to monitor infections and adverse events was already in 1991 when Anne Gardulf and co-

workers published the first report on subcutaneous IgG replacement therapy for IgA- and IgG-deficiencies (Gardulf et al, Lancet, 1991). A similar questionnaire was recently used in a clinical study on asthma in children (Asthma Control Questionnaire-score, Holbrook et al, JAMA, 2012). In our current study, all patients had to provide information on the number of days with infection (during the year prior to inclusion) to the study physician at the time of inclusion and only those presenting more than 42 days with symptoms of respiratory infection were included. Thus, we believe that we used a robust instrument for evaluation of disease burden that has been used in the field for many years. It is very practical and useful way of assessing infectious symptoms and antibiotic consumption. In addition, on each visit the patient was specifically asked only to register symptoms directly related to an ongoing respiratory tract infection, including malaise. Hence, all points registered are directly representative of clinical symptoms of infections from the respiratory tract. This information is mentioned in the manuscript (page 8, final paragraph).

--What is the clinical significance of the differences observed in this population?

To determine the effect size of the study, we estimated that a reduction of infectious score with 30% (210 points => 140 points) would be clinically significant. However, this estimation was rather arbitrary and based on our clinical experience but does not exclude that a difference lower than 30% could be relevant. In fact, the intervention group showed an absolute reduction of 47 points, which was lower than the expected reduction of 70 points. The effect in the adjusted analysis was 23% reduction. Although the effect was statistically significant, it could be argued that this is not clinically relevant since it is lower than the pre-defined cut off value of 30%. Nevertheless, we believe that the 47 points reduction on average for patients in the intervention group can be clinically relevant, since it translates to 47 days with symptoms from the nose, ear, sinuses or respiratory tract (47 days x 1 point). It can also be interpreted as 15 days with cough, a runny nose and antibiotics (15 days x 3 points). (This is explained in the last part of the discussion section, already included in the original version, page 20, lines 1-10.)

--For both the inclusion criteria and composite outcome, how are allergic symptoms (which can include many symptoms from the respiratory tract), distinguished from infection. How about chronic symptoms that are also not infectious (e.g, reflux disease, chronic asthma, allergic rhinitis, etc).

Upon inclusion all participants were specifically instructed to only register symptoms related to infections. During all visits to the study site (inclusion, 6 months and 12 months) the patients met with a physician (PB or ACN) and doctor and patient reviewed the diaries together. Symptoms that were totally unrelated to infectious symptoms were omitted. We believe that this follow-up and detailed interview with the patients helped to get a better estimation of true infectious symptoms.

--This is a select group of patients and the findings very well may not generalize. This should be emphasized in the conclusions.

This point is important and there is a clear statement in the discussion part that the results from this study cannot be extrapolated to the general population. However, as written out, the results provide support for further studies among patients with frequent respiratory tract infections and a high consumption of antibiotics. (page 17, first paragraph and page 19, final paragraph).

--Unclear when the 25OHD levels were performed (i.e., if during the trial, how were those that assessed outcomes blinded from the results)?

All safety parameters (clinical chemistry and S-25-OHD levels) were reported to the monitor, who kept these files separately and blinded to the study team (PI, doctors, nurses). When the study was completed the data was reported to the Principal Investigator (JA) and analysed by the study team.

--I would like to see more information about the analysis of the primary outcome in the main manuscript.

We chose to describe most of the statistical methods in the supplementary section only due to restrictions regarding the length of the manuscript, but fully agree that more information would be valuable in the main manuscript. To comply with the request, we have extended the methods section of the main manuscript with detailed information about the main analysis (pages 9-11).

--Unclear whether intention to treat analysis was performed and for which outcomes, given that there were post-randomization exclusions (were those subjects actually included in the analysis)? If all n=70 per group were included, how were missing data handled?

According to the pre-specified analysis plan, analyses were to be performed per-protocol. However, we have extended the plan with an intention-to-treat analysis of the main outcome (total infectious score) with imputation of missing values. As requested by reviewer 3, the original imputation method has now been substituted with a more advanced and well-documented multiple imputation method (see below). Notably, both imputation methods produce ITT effect estimates very similar to those derived from the PP analysis.

--Since AMPs and bacterial cultures were performed at set intervals (unknown whether sick or well), unclear on the value/interpretation of these results are without a consistent protocol for measurement. Also, unclear why the placebo group had nearly double the number of cultures. What was the standardization of measurement for these items?

Samples for bacterial culture were collected according to two principles: i) every 5th participant (n=15) was randomized to leave NPH-aspirate for AMP-analysis. Since the bacterial flora in the nasopharyngeal tract has a major impact on AMP-levels in nasal fluid, a bacterial culture was also taken (Cederlund et al, PLoS One, 2011); ii) all participants were instructed to leave a bacterial culture when they experienced symptoms of infection. Thus, the first procedure was evenly distributed between the groups due to randomization but the second procedure (to leave a culture when having symptoms) was left to the patients to decide.

--Why then did the placebo group have almost twice the amounts of cultures?

We believe that the lower frequency in the intervention group was due to the clinical effect of the vitamin D treatment (fewer infectious episodes prompting sampling for bacterial culture), but of course other mechanisms may be involved. Theoretically, any number of patient-level factors could have influenced the propensity to leave a bacterial culture, but due to the double-blind study design this is unlikely to account for the observed between-group differences. We discuss possible mechanisms for vitamin D-mediated effects on mucosal immunity as well as other causes of our observations on page 19, first paragraph).

--The CONSORT checklist has an item 95%CIs be provided for each analysis. This is not true for most of the data presented.

The 95% CI has now been added for all data generated by linear regression. However, for data generated using non-parametric analysis or contingency (Fisher's exacts test) we decided to only write the absolute difference and the corresponding p-value.

--The CONSORT checklist has an item that the number of participants (denominator) be included for each analysis-- this has not been done and it is unclear what the denominator is for the analyses (if

true intention to treat, should be n=70 per group for each analysis).

Thank you for pointing this out. Numbers of patients included in the analyses have now been added in the results section and in the tables.

--In Tables where p-values were provided, would provide the actual value, instead of "NS".

We have added actual p-values also for non-significant results.

--The first supplementatal figure is missing the "n" for the "Completed Vitamin" item.

This point presumably refers to Figure 1, where some information was missing. A new Figure 1 (flowchart) has been made.

Reviewer 3:

--The outcome measure is a composite measure and it is not stated if this has been validated or any training/education that the patients need to complete this questionnaire. How well are the patients able to identify symptoms related to ongoing respiratory tract infections?

A similar diary has been used to monitor these types of patients for more than 20 years. The first use of a diary to monitor infections and adverse events was already in 1991 when Anne Gardulf and co-workers published the first report on subcutaneous IgG replacement therapy for IgA- and IgG-deficiencies (Gardulf et al, Lancet, 1991). A similar questionnaire was recently used in a clinical study on asthma in children (Asthma Control Questionnaire-score, Holbrook et al, JAMA, 2012). In our current study, all patients had to provide information on the number of days with infection to the study physician at the time of inclusion and only those presenting more than 42 days with symptoms of respiratory infection were included. Thus, we believe that we used a robust instrument for evaluation of disease burden that has been used in the field for many years. It is very practical and useful way of assessing infectious symptoms and antibiotic consumption. In addition, on each visit the patient was specifically asked only to register symptoms directly connected to an ongoing respiratory tract infection, including malaise. Thus, all points registered are directly representative of clinical symptoms of infections from the respiratory tract. This information is now mentioned in the manuscript (page 9, final paragraph).

--The sample size is based on using a p-value of 0.02, this is unusual and needs justification. They should also use this in the paper by reporting 98% confidence intervals and use this significance level for testing. If they do then there discussion and results of the primary outcome are incorrect since it does not meet the 0.02 cut-point. The sample size calculation does not specify the standard deviation used nor does it justify the effect size they are looking for or the drop-out rate.

We totally agree with the reviewer that our sample size calculation can be criticized. We had great difficulties with the estimation of sample and effect size, since no similar studies had been done prior to ours. Thus, we had to define arbitrary values for effect size and standard deviation. The effect size was estimated as a reduction from 6 weeks with full symptoms to 4 weeks (42 days x 5 points=210 points to 28 days x 5 points = 140 points) with an estimated SD of 3 weeks (21 days x 5 points = 105 points). This resulted in a sample population of 60 patients per arm. To adjust for a dropout of approximately 10% we added 10 patients per group. Thus, the study included 70 patients per arm in the final design. The choice of p=0.02 was unfortunate but we aimed for as large study group as possible, given the assumed in-data. The intention all along was to use the classical p=0.05 in the final analysis. The correct way to reach the same sample size would of course have been to increase

the targeted power of the study, but we instead reduced the p-value as a means of asserting a sample size large enough to produce results significant at the p=0.05 level. In addition, we assumed that we could use a standard t-test for the final analysis. However, some of the data were non-normally distributed and we also wanted to adjust for a number of important factors (age, sex, smoking etc) and thus chose to use linear regression. However, we firmly believe that although our power calculation and pre-defined statistical analysis plan was sub-optimal it does not affect our data analysis nor the main conclusions of the study.

--In the statistical analysis they adjust for age, gender, smoking status, type of immune deficiency and co-morbidities. They do not justify these choices or state if they are based on previous research suggesting a link with the outcome measure or imbalance at baseline (which would be a poor justification).

We chose to correct for a number of factors, which we consider to be relevant for the outcome. For example, we know from our clinical practise that older people and smokers get more infections than younger individuals. There are a number of reports showing that these factors affect susceptibility to infections. Moreover, women tend to have more infections and more concomitant diseases than men at our centre and it was thus interesting to adjust for gender. Since we know that both the underlying immunological diagnosis and other disease may influence the infection rate, we adjusted for these factors as well.

--The results are credible and do relate to the research question. However, they do not appear to be consistent with the sample size calculation. Firstly, they use a different level of significance (5% in the analysis, 2% in the sample size calculation).

Please, find a detailed explanation for the use of p=0.02 above.
Secondly, they appear to consider a 23% reduction important in the analysis whereas the sample size calculation implies that the smallest clinically important difference is 30%. Since we could not find any similar study to base our assumption of effect size on, we defined an arbitrary effect size based on our clinical experience. We reasoned that if we could reduce the number of days with full symptoms from 42 days down to 28 days (translated to points: 210 points => 140 points) for the average patient, this would be clinically significant. However, this assumption does not exclude that effects lower than 30% could be clinically relevant as well. In fact, the intervention group showed an absolute reduction of 47 points, which was lower than the expected reduction of 70 points. The effect in the adjusted analysis was 23% reduction. It could of course be argued that this is not clinically relevant since it is lower than the pre-defined cut off value of 30%. Nevertheless, we believe that the 47 points reduction on average for patients in the intervention group can be clinically relevant, since it translates to 47 days with symptoms from the nose, ear, sinuses or respiratory tract. It can also be interpreted as 15 days with cough, a runny nose and antibiotics. This line of reasoning is written out in the discussion part (page 21, final paragraph).

--Additionally, the unadjusted analysis is based on a Mann-Whitney test whereas the adjusted analysis is based on a regression model with a log-transformed outcome....this makes is impossible to compare the adjusted effect and the unadjusted effect to see if the adjustment made any difference to the effect size.

We agree that the choice of the Mann-Whitney U test for the unadjusted and linear regression on log-transformed values for the adjusted analyses make the effect of the adjustment difficult to quantify. Hence, we have now substituted the non-parametric test with a simple regression (of log-transformed scores). This makes the results of the adjusted and unadjusted analyses easy to compare, and the change of analysis method had very little impact on the outcome. For example the p-value of the main analysis (total score over full study period) changed minimally, from 0.023 to 0.024.

--The analysis is also difficult to follow as it is unclear what population the analysis refers to. Normally, it would be the intention-to-treat analysis which would be the main analysis.

According to the original analysis plan, the study was to be analysed per protocol (this is why no further outcome data was registered in patients dropping out of the study). Rather than changing the plan post hoc, we have chosen to adhere to the original intention. However, along the way we have come to realise that it may be insufficient to present the study results without an analysis according to intention-to-treat. Hence, we have added an ITT analysis of the primary outcome as a secondary analysis. Notably, the ITT analysis using the standard imputation method for missing data, resulted in almost the same result as the per-protocol analysis.

--It is also unclear to me why the authors have chosen to use an ad-hoc imputation method rather than the standard imputation approaches. The statistical properties of this imputation process are not known (for example how does it allow for uncertainity in the imputation process) and hence it shouldn't be used. The statement that it is conservative seems to be made without justification and cannot be true in general, for example if those who dropped out of the placebo group were better than those who remained in the trial then replacing them by the mean of those who remained in the trial would make the placebo group appear worse than they were.

The main reason for using a non-standard imputation method was that we wanted to make use of the observed data in patients with incomplete diaries (by imputing only those days where data was missing). Since many infections are not evenly distributed over the year, LOCF did not seem like an ideal choice, and we therefore decided on a simple ad-hoc method of day-by-day imputation. However, we agree that the choice of a well-documented imputation method taking patient characteristics into account and providing correct estimates of variability may be preferable. Hence, we now use multiple imputation with predictive mean matching, with similar results.

--The OR for antibiotic use was 0.365 which is losely interpreted to be a "63.5% reduction" it should be clearer that this is a reduction in the odds of use not a 63.5% reduction in use of antibiotics.

This is an important point and has been corrected in the text (page 13, line 12; page 17, line 20; page 21, line 16).

--The subgroup analysis is a little unclear if this is done on the log-transformed outcome or the untransformed. It is clear from the supplementary material what was done, but it is unclear in the main paper.

This has been clarified in the revised version.

--Overall, I think that this is an interesting study which was well conducted but I think that the analysis of the data can be made much clearer.

We thank the reviewer for all comments and suggestions. Overall, we have tried to explain better how the data was analysed.

Some minor points:

--i) you use "," rather than "." in many places.
We assume that this point refers to the use of ”,” in decimal numbers. This has been corrected all through the text.

--ii) p-values should not be reported in Table 1.
Table 1 is corrected (p-values are removed) and expanded with data on concomitant lung disease (as suggested by reviewer 1).

--iii) actual p-values should be given rather than "n.s.".
This is now corrected all through the text.

VERSION 2 – REVIEW

REVIEWER	Allan Clark University of East Anglia
REVIEW RETURNED	09-Oct-2012

THE STUDY	I'm pleased that the authors have address my minor comments. However, the authors have not adequately addressed my previous comment about using p=0.02 in the design. This choice is part of the design of the study and cannot be simply ignored when it comes to the analysis. I am concerned that in the response the authors claim to have "aimed for as large a study group as possible" the purpose of sample size calculations is to ensure that the number of participants is neither too small or too large but that it is appropriate to answer the study questions in an ethical way.
RESULTS & CONCLUSIONS	The size of the effect found by the authors is smaller than what they considered clinically meaningful in the design and is not significant at there pre-specified level of significance (p=0.02). There are two issues a) the size of the effect. I agree that the authors have discussed the interpretation of the difference and why they consider it clinically meaningful, but I would prefer a direct comment on why this has changed since the design of the study. b) the choice of significance level. The planned level of significance testing was p=0.02 which would make the results not significant. I think that the authors have been unfortunate in this respect but I feel that have done a good job in presenting the actual p-value so that readers can judge for themselves; however they should carry through the design of the study and discuss the results according to the threshold they set. They could of course discuss at the traditional 5% level as well.

VERSION 2 – AUTHOR RESPONSE

Q: I'm pleased that the authors have addressed my minor comments. However, the authors have not adequately addressed my previous comment about using p=0.02 in the design. This choice is part of the design of the study and cannot be simply ignored when it comes to the analysis. I am concerned that in the response the authors claim to have "aimed for as large a study group as possible" the purpose of sample size calculations is to ensure that the number of participants is neither too small nor too large but that it is appropriate to answer the study questions in an ethical way.

Answer: We thank the reviewer for pointing this out. Of course, the power calculation is used to obtain an optimal study group. However, in the context of our vitamin D trial the power calculation was complicated by a great uncertainty regarding the parameters involved. When accounting for this uncertainty, we had to make a choice between the risk of ending up with a too small sample and the

the risk of performing an unnecessarily large study. Risking a too small study population, would have been unethical and problematic in the sense that we probably would not have reached a meaningful conclusion. On the other hand, there is a problem with designing a large study since you then include more patients than is needed in order to answer the research question. Given these two options and the presumably non-hazardous interventions involved, we chose the latter, i.e to introduce a margin of error by designing a slightly larger study. Thus, a more stringent p-value of 0.02 was used for the power calculation even though we had the full intention to use p=0.05 in our final analyses of both primary and secondary endpoints. The second choice – a slightly larger study - was particularly important given the problems of defining a clinically meaningful effect and the expected drop-out rate during a one year study period. Importantly, we believe that our study design has been ethical and that our study size of n=140 was appropriate for our research question.

Q: The size of the effect found by the authors is smaller than what they considered clinically meaningful in the design and is not significant at there pre-specified level of significance (p=0.02). There are two issues

a) the size of the effect. I agree that the authors have discussed the interpretation of the difference and why they consider it clinically meaningful, but I would prefer a direct comment on why this has changed since the design of the study.

Answer: Again, we thank the reviewer for helping us to clarify this crucial point. We have now extended the discussion part (page 21, final paragraph) on the discrepancy between the predefined level of what was tentatively considered to be a clinically meaningful effect (30%) and the actual observation (23%). This will now help the reader to better interpret the main results of the study.

b) the choice of significance level. The planned level of significance testing was p=0.02 which would make the results not significant. I think that the authors have been unfortunate in this respect but I feel that have done a good job in presenting the actual p-value so that readers can judge for themselves; however they should carry through the design of the study and discuss the results according to the threshold they set. They could of course discuss at the traditional 5% level as well.

Answer: Please, find a detailed discussion on the rationale for p=0.02 above. We fully agree with the reviewer on this issue and in the revised manuscript we have tried to make it very clear to the reader how we have reasoned regarding significance levels. We have made several changes in the text:

First and most importantly, it is important to point out that we did not intend to use the significance level of p=0.02 for the statistical analysis of the primary endpoint, since this would deviate from the convention in the field of clinical trials. In fact the significance level of p=0.02 was only used for the power-calculation and was not a part of the prespecified analysis plan for the primary endpoint. Unfortunately, this was not explicitly stated in the full protocol from 2009 but as written now it should be very clear to the reader how we have reasoned all along the project (page 7, final paragraph).

Second, we have moved information on the power calculation to the main text (page 7) from the supplementary section, since it is central to the understanding of the main results.

Thirdly, we now present this discrepancy as a weakness of the study as a whole in the discussion part (page 18, lines 2-7). In addition, the reader can also find the actual p-values all along the text, which will help in the interpretation of the results.

Taken together, we think that the important point concerning the power calculation is now properly introduced (M&M-section) and discussed (discussion-section). In addition, all p-values are written out.

Thus, the reader should be able to fully understand what has been done, why it has been done and finally, how to interpret the main outcome of the study.

VERSION 3 - REVIEW

REVIEWER	Allan clark University of East Anglia
REVIEW RETURNED	23-Oct-2012

THE STUDY	The authors have not adequately addressed my concern about the difference between the sample size and the analysis in terms of the significance level used. The authors say they use p=0.02 in order to "ensure that sufficient number of patients were recruited in order to avoid a type II error". This is methodologically incorrect, in order to reduce the risk of a type II error you would increase the power of the study you do not reduce the risk of a type I error. Sample size calculations are complex to undertake and cannot be done by simlpy changing the parameters to ensure that you get a number which you think is reasonable, each parameter value used must be justified (or at least commonly used) and has consequences for the interpretation of the results.