Subject: Our vitamin C and asthma paper in CTA // Fwd: RE: Data
From: "Hemilä Harri" <harri.hemila@helsinki.fi>
Date: 3.1.2012 11:10
To: "Lotvall, Jan" <jan.lotvall@gu.se>, "Bell, Emma" <Emma.Bell@biomedcentral.com>, "Phelps, Lisa" <Lisa.Phelps@biomedcentral.com>
CC: "Al-Biltagi, Mohammed" <mbelrem@hotmail.com>

Dear Dr. Jan Lötval,

Re: my paper in CTA

http://www.ctajournal.com/content/1/1/9

I received the corrected data set from Dr. Biltagi on 19 Dec 2011 (see his EXCEL file "Data" as an attachment).

I have been analyzing and thinking the consistency of the corrected data set for over a week.

Please, find attached my EXCEL file ("--Problems--"), in which I show my calculations of the corrected data set.

There are numerous severe problems with the corrected data set:

1

In the upper half (first 30 children) of the corrected data set (19 Dec 2011), the range for weight is from 23 to 36 kg.

In the lower half (last 30 children) of the corrected data set, there are several children outside the weight-range of the upper half.

We should not expect identical range in the upper and the lower halves, but several children in the lower half are far from the range of the upper half:

14, 15, 47, 61, 66, 75, and 77 kg

The lowest and the highest weights in the lower half (last 30 children) are far from the usual weights of 7 to 10 year old children.

According to WHO graphs, such weights in the age range 7 to 10 years are essentially impossible in a set of 30 children

http://www.who.int/growthref/who2007_weight_for_age/en/index.html

http://www.who.int/entity/growthref/cht_wfa_boys_perc_5_10years.pdf

http://www.who.int/entity/growthref/cht_wfa_girls_perc_5_10years.pdf

2

We can adjust weight for height and analyze BMI values, which gives an independent look at the data.

In the upper half of the corrected data set, the range for BMI is from 15 to 22 BMI-units.

In the lower half of the corrected data set, there are several children outside the BMI-range of the upper half.

We should not expect identical range in the upper and the lower halves, but several children in the lower half are far from the range of the upper half:

11, 30, 33, 39, and 41 BMI-units.

According to WHO graphs, such BMI values in the age range 7 to 10 years are essentially impossible in a set of 30 children

http://www.who.int/growthref/who2007_bmi_for_age/en/index.html

http://www.who.int/entity/growthref/cht_bmifa_boys_perc_5_19years.pdf

http://www.who.int/entity/growthref/cht_bmifa_girls_perc_5_19years.pdf

These children with 11, 30, 33, 39, and 41 BMI-units are also far from the usual range of BMI-values for Egyptian boys of age range 7 to 10:

http://3.bp.blogspot.com/_CYYc5niOLdc/STGA-uBNK7I/AAAAAAAAAB0/xiKXw-NCbeM/s1600-h/Picture6.jpg

http://dempuegypt.blogspot.com/

3

On the basis of published methods, the children should be homogeneous so that we should not expect systematic baseline differences between the upper and the lower halves.

However, the average of BMI is 3 units higher in the lower half and the difference between the upper and the lower halves is highly significant in the t-test (P=0.007).

Furthermore, the SD for BMI in the lower half is 4.5 times as great as the SD in the upper half, which also is highly puzzling (see also problem 2 above). How can the first 30 children be so different from the last 30 children (when their ages are identical, see problem 4 below)?

4

The upper (first 30) and the lower (last 30) halves of the corrected EXCEL file (19 Dec 2011) have exactly the same mean and SD for three variables:

age, ACT after Zn phase, FEV1 after Zn phase.

As shown by sorting, the observations for these three variables are identical in the lower halves of the old and the corrected EXCEL versions: only the order is different.

In the old versions, the upper halves were copy-pasted to the lower halves keeping the order identical.

In the corrected version, the upper halves are copied to the lower halves, but the data is shuffled to make the order different.

Whatever is the order in the lower half, it is extremely unlikely (impossible) that there are, purely by chance, 30 identical ages and 30 identical zinc period ACT observations and 30 identical zinc period FEV1 observations in the upper and the lower halves of the data set.

5

The upper (30) and the lower (30) halves of the corrected EXCEL file have exactly the same means for:

ACT and FEV1 after vitamin C phase.

Given the inaccuracy of FEV1 (=high SE-value =32), it is extremely unlikely (impossible) that, purely by chance, the upper and the lower halves might have exactly the same means with the accuracy of the first decimal: 1446.2.

For FEV1, the data are identical in the lower halves (last 30 children) of the old and the corrected EXCEL versions except for 2 children. The order is different between the old and the new data set versions, as for the three variables listed above. Thus, the data of the lower half has been shuffled also in this case.

Given the SE value (=0.36) for the ACT, it is unlikely that the upper and the lower halves of the ACT values for vitamin C phase might have exactly the same average with the accuracy of the first decimal: 19.6. This identity in mean is not as extremely improbable as the identity of the FEV1 means above, but even this is highly unlikely.

6

Previously I did not compare the up and low halves of “dampness” variable.

In this corrected data set, there is a high correlation between the up and low halves (r=0.93), so that only 3 children in the lower half have different data compared with the series in the upper half.

Thus, the dampness data has also been copy-pasted from the upper to the lower half in the same order with the exception of the 3 children.

The lower half of dampness has not been shuffled in the corrected EXCEL file apparently because I did not point out previously the nearly exact duplication of the dampness data.

7

In addition to the means and SD:s described above, the validity of a data set can be analyzed by calculating correlations between different variables.

There are several correlations that we expect on the basis of biology.

For example, we expect that on average older children are taller and heavier.

Thus, we expect positive correlation between age and height/weight.

However, this correlation is seen only in the lower half of the corrected EXCEL:

correlation between age and height is

r = 0.03 in the upper half (first 30 children)

r = 0.45 in the lower half (last 30 children)

correlation between age and weight is

r = 0.09 in the upper half (first 30 children)

r = 0.71 in the lower half (last 30 children)

8

We expect older children to have greater FEV1 values.

However, out of three calculations that I carried out, all the strong correlations were restricted to the upper half, and one lower half comparison found a negative correlation between age and FEV1 (after placebo period).

Although for some very strange group of children there might not be a positive correlation between age and FEV1, I cannot imagine any explanations for the extreme divergence in the correlations between the upper half (first 30 children) and the lower half (last 30 children):

correlation between age and FEV1 after zinc period is

r = 0.89 in the upper half (first 30 children)

r = 0.24 in the lower half (last 30 children)

correlation between age and FEV1 after vitamin C period is

r = 0.93 in the upper half (first 30 children)

r = 0.13 in the lower half (last 30 children)

correlation between age and FEV1 after placebo period is

r = 0.96 in the upper half (first 30 children)

r = -0.17 in the lower half (last 30 children)

9

We expect taller children to have greater FEV1 values.

However:

correlation between height and FEV1 after placebo is

r = 0.06 in the upper half (first 30 children)

r = -0.19 in the lower half (last 30 children)

10

In the old data set (vitamin C data one year ago), the third, fourth and fifth (etc) children had FEV1 values after placebo phase 1020, 1090, 1060 (etc). In this corrected data set (19 Dec 2011), the same lines have data 1022, 1093, 1064.

Given that FEV1 does not have greater accuracy than some 1%, why do these new corrected lines have greater accuracy than the old vitamin C data set which I received one year ago?

For the FEV1 value after the vitamin C period, the accuracy is in the units of ten milliliters so that the same lines have values 1330, 1400, 1380, which seems reasonable accuracy to record FEV1 measurements.

Thus, why and where does the improved accuracy for the FEV1 placebo values come from for the corrected data set? These changes are small but highly puzzling.

In summary:

Given the description of the methods in the Acta Pediatrica (2009) paper by Dr. Biltagi, it is not possible to get these kinds of findings:

There are substantial differences between the upper and lower halves that are impossible given the homogeneity of the group of the children (problems 1 to 3)

There are very close similarities that are impossible given the random variation which is always present in experiments (problems 4 to 6)

There are severe inconsistencies in the correlations between biologically relevant variables (problems 7 to 9)

****

When I received the Egyptian vitamin C data set over one year ago, I found some errors in the data set, but they were not fundamental.

Weigh was 127 kg for two children and height was 17 cm for two children.

I classified those errors ordinary sloppiness. Of course, a careful researcher looks through the weight and height columns to check that none of the values are impossible, and he or she can use the functions available in EXCEL such as min and max to search for possible errors in coding (i.e. when checking the transfer of data from papers to EXCEL). Etc.

I pointed out those errors to Dr. Biltagi and he corrected them.

Although I understood from those errors that Dr. Biltagi and his colleagues were not careful, I did not consider that such sloppiness was an indication to suspect the validity of the entire data set.

When I received the zinc data set and found the exact duplication of the upper half to the lower half for a few variables and a nearly exact duplication for several other variables (30 Nov 2011 forward), I considered that such a great error could result either from extra-ordinary sloppiness or from fabrication.

By extraordinary sloppiness I was thinking, for example, that there may be a laboratory book containing the recorded measurements, where the original data had been collected. Children from 1 to 30 might be on one page and children from 31 to 60 might be on the second page. In such a case it might be possible to mess up the two pages and write the same data sequence from one page twice to the EXCEL file and that kind of sloppiness might have explained the duplications.

With such kinds of possibilities in my mind, I asked Dr. Biltagi to check and correct the EXCEL file against their original paper records. I assumed that they had the original records somewhere. Given that the study was published in 2009, the original data should be stored for several years and should be still available.

When I found the duplication problem, I tended to favor the notion of extra-ordinary sloppiness, since I could not believe that anyone would invent data simply by copy-pasting the same series in the same order.

Dr. Biltagi must understand that when there starts to be suspicion about the validity of their data set, he must try to find out what caused the problems. However, he has not given any explanations, not even speculative explanations, about what might have caused the duplication. Furthermore, Dr. Biltagi does not describe in any way where he got the data for the lower part of the “corrected” data set. For example, how is it possible that in the corrected data set the results in the upper half are identical with the lower half for three variables (Problem 4) although the purpose was to correct such an obvious error. Shuffling the data is not a correction.

I encouraged Dr. Biltagi to ask some of his colleagues to carry out the data checking independently of him. The original Acta Pediatrica (2009) paper had five authors, and the other four authors should be concerned about the validity of the data. If there are several persons

independently checking the consistency of the original paper records against the corrected EXCEL file, that would substantially increase the credibility of the corrected data set.

However, Dr. Biltagi simply wrote to me that he had sent the corrected data set to the other authors, but none of the other authors has confirmed to me that they have independently checked the corrected EXCEL file against the original paper records (see an attached Biltagi's email on this issue).

Thus, one month ago I reasoned that my CTA paper might be saved:

1) if Dr. Biltagi would send me a valid corrected data set

and

2) if some of his collaborators would independently confirm that they have also checked the corrected data set against the original results.

Neither of these two issues materialized.

It seems probable to me that the Egyptian group did an experimental study. However, the "corrected" data set (Dec 19, 2011) cannot be the results of their study. It is not possible to get that kind of data set from a real experiment (see problems 1 to 10). Thus, possibly the original observations have been lost and there is only computer generated junk which has no relation to the experiments. Possibly some parts of the EXCEL data set have been generated by their original experiment, but we do not know which parts. Furthermore, large parts of the data are not aligned properly as shown, for example, by the inconsistency in the correlations (problems 7 to 9).

I do not assume that Dr. Biltagi and his colleagues have invented the study results, since they would not have given the data to me if they knew that it was fundamentally flawed. It seems probable that because of extra-ordinary sloppiness, the original experimental results have simply vanished and there is remaining a data set which I think we can classify fabricated if that is described as the "results" of the study.

When I found the duplication problem one month ago (Nov 30, 2011), I assumed that the upper part of the EXCEL file might be valid, so that only the lower part might be flawed. However, when I now took a closer look at the data set, I found that there is no correlation between age and height/weight in the upper half. Since biology universally causes a correlation between age and height/weight in children, either age or height/weight or all three of them must be fabricated in the upper half. Therefore, I lost my trust also on the upper half. Thus, the effect of vitamin C on asthma might depend on age, but nothing can be concluded on that issue from the Biltagi data set.

I am sorry for submitting and publishing a paper which I now understand is based on fabricated data. However, when I was doing the statistical analysis, I did not have reason to suspect the data. Furthermore, had I calculated the correlations between age and height/weight at the whole data set level (all 60 children), the correlation would have seemed reasonable. It was the comparison of the upper half against the lower half that discredited the

data set. That kind of problem cannot be found in ordinary data checking. It was the finding that half of the data points were missing from one figure which led me to systematically examine the inconsistencies between the upper and lower halves.

The goal of science is to get reliable information and increase our understanding. However, producing valid results is not the only way forward, but sorting out which results cannot be trusted at all helps us also forward. Without my analysis of the Biltagi's data set, many people would keep on trusting on their 2009 paper. In this respect my adventure with the Egyptian data set has also benefits as we can now firmly conclude that the 2009 paper does not teach us anything about asthma treatments. Unfortunately, publishing my paper in CTA is a rather high price for finding the lack of validity of the Acta Pediatrica 2009 paper. Given that Dr. Biltagi has given his word that the Dec 19 EXCEL file is corrected ("I spent a lot of effort to be sure no mistakes." see below), I cannot see that any parts of the 2009 paper can be saved.

I will inform the editor of Acta Pediatrica that their 2009 paper is based on seriously flawed data.

Could you please instruct me how I can retract my CTA paper.

How much is it reasonable to use words for describing the identification and evidence that the data set is flawed.

Is it possible to attach an EXCEL file to the retraction so that the readers can by their own eyes see the main problems.

Happy New Year otherwise (... not so happy with this specific issue...)

Yours sincerely,

Harri Hemilä
Helsinki, Finland

****

This below is Biltagi's email which I received Dec 19 without any description where the corrected data came from.

****

----- Forwarded message from mbelrem@hotmail.com -----
Date: Mon, 19 Dec 2011 20:39:36 +0200
From: "Mohammed Al-Biltagi" <mbelrem@hotmail.com>
Subject: RE: Data
To: "Harri Hemila" <harri.hemila@helsinki.fi>

Dear Dr Harri

Good day

This file contains the revised data. I spent a lot of effort to be sure no mistakes.

Thanks
Mohammed

----- End forwarded message -----

Attachments:

Data.xls	33.0 KB
Biltagi_Email_2011_12_23.pdf	66.8 KB
Biltagi_NewData_Problems_2012_01_03.xls	50.0 KB