Department of Surgery and Cancer
Division of Cancer
Imperial College London

Imperial College
London

Institute of Reproductive and Developmental Biology
(IRDB)
Hammersmith Hospital Campus
Du Cane Road
London
W12 0NN

Tel:+44 (0) 207 594 2104 Fax: +44 (0) 207 594 2184

ilpo.huhtaniemi@imperial.ac.uk
www.imperial.ac.uk
www1.imperial.ac.uk/medicine/about/institutes/irdb
Professor Ilpo Huhtaniemi MD PhD MDhc FMedSci
Professor Emeritus of Reproductive Endocrinology

April 13th, 2016

VTT Technical Research Centre of Finland Ltd (c/o Laura Puronen)

Dear Sir/Madam:

I thank VTT for inviting me to provide "*an informed and independent second opinion as to whether there is unequivocal proof of fraud (data tampering, falsification, or erroneous reporting), excessive interpretation of results or unrealistic conclusions*" in the publication Oresic et al. (J Exp Med 2008;205:2975-2984). As instructed I will limit my evaluation "*to this scientific article alone*".

As evaluator I have some strengths and weaknesses. On the strength side I have over 45-year experience in biomedical research, as author of about 650 scientific publications in the field of experimental and clinical endocrinology, as Editor-in-Chief for 18 years of the journal Molecular and Cellular Endocrinology, publishing research very similar to the JEM article, and as evaluator of grant applications and scientific manuscripts in the field of translational endocrine research. I consider myself unbiased and impartial with respect to this task. I have had no personal or professional contacts with prof. Oresic. I am co-author in two publications (2002 and 2003) with one of the co-authors, prof. Mikael Knip. My laboratory carried out hormone measurements for these studies, but I had no contact with Knip, and was not aware of his involvement, during conduction of these studies. My limitations are that I have no special expertise in statistics and in technical details of metabolomics studies, but I have familiarized myself with the techniques and understand their strengths and weaknesses. The statistical evaluation of the study has already been done twice (Tale Oy and 4Pharma Ltd), and I trust this aspect has been properly handled.

General comments

The JEM paper is based on analyses of serum samples collected in two large Finnish prospective studies on childhood diabetes (DIPP and STRIP). I assume the original goal of these studies was not to carry out metabolomics studies and the analyses in JEM were carried out on 'convenience samples' collected for other purposes and not necessarily designed and planned for the needs of metabolomic profiling. One potential caveat is that blood samples were collected without fasting and there was no standardization of diets, which may contribute to the variability of the results. However, fasting status has been reported only to contribute very little to the variability in metabolomics studies (Sampson et al. Cancer Epidemiol Biomerkers Prev 2101;22:631-640). A major weakness of this study is its small sample size. According to Nicholson et al. (Mol Syst Biol 2011; doi:10.1038/msb.2011.57): *"sample sizes of 2000-5000 should allow reliable identification of disease-predictive metabolite concentrations explaining 5–10% of disease risk, while greater sample sizes of 5000–20 000 would be required to identify metabolite concentrations explaining 1–2% of disease risk*." On this basis the study is much too small to allow strong conclusions. It can mostly be considered a hypothesis generating exercise. At the time of publication this article metabolomics was in its infancy, and as is common in science the publication criteria tighten with time. This paper may have been up-to-date in 2008, but not any more. I have taken this into account in my evaluation.

A quality journal, such as JEM, is expected to do proper quality assessment by peer review of the data it publishes, including the statistical analysis and whether all interpretations are backed by statistical significance. In this respect the journal has been surprisingly relaxed by allowing statements of differences with no statistical proof and allowing quite far-fetched speculations in discussion. It is interesting to speculate why JEM accepted this study despite its weaknesses. If a study is considered to be potentially highly interesting, predicting a paradigm shift in the field, it is understandable that small technical faults are overlooked, if the main message is considered really important. This is likely the reason why JEM published the study.

Reviewer comments of JEM

In line with the above, it was informative to read the JEM referee reports of the study, which sharply pointed out the main weaknesses of the study as well as its potential strengths and long-tern repercussions.

Referee 1 stated that '*the results demonstrated some promising results in terms of identifying biochemical changes that precede type 1 diabetes*', but '*for many metabolites there appears to be incomplete separation between the groups*', making their clinical significance questionable. Analysis of a larger material for fewer analytes was considered better than a large number of analyses from a relatively small sample. The referee wondered the representativeness of Figure 2 for the whole material. The mechanistic link between the findings and diabetes, as concerns Figure 6, was considered tenuous and too speculative. The referee concludes that '*there is an interesting story in this study*.' In specific comments the organisation of data presentation was criticized. About statistical analysis the lack of correction for multiple comparisons and statements of claims of changes based on nonsignificant changes were reprimanded, as well as confusing statements whether there is a change or not in particular metabolites, and whether the authors were confident of them.

The comments of *Referee 2* were more general. He/she considered the study '*truly ground-breaking*' and '*if confirmed, the metabolic alterations will shake diabetes research out of its present line*.' Furthermore, it was stated that '*The authors have carefully collected the metabolic alterations, and their observation, if confirmed, may have major implications in the way we understand the insult to beta-cells*.' The lack of mechanistic explanation was noticed, but not criticised. Finally, it was noted that '*the only risk is an artifact'* in the form of '*stratification of the progressor group*'.

The positive tone of referee 2 in particular, praising the ground-braking potential of the study, yet admitting the risk of artifact, may have been the reason why this technically imperfect study was accepted to a top-tier journal. The journal apparently decided to take a calculated risk of publishing something that would be on somewhat shaky ground, but would represent a paradigm shift if verified by later studies. Importantly, the referee critique did not point our any suspicions for any type of scientific misconduct.

Opinions in articles citing the JEM study

Another source of information about the quality of the paper is to scrutinize how the scientific community has received it. The JEM paper has been cited since its appearance 168 times. Most of the studies reported the finding of metabolomic changes briefly and without qualitative assessments. The only qualitative comments I found in the citations were the following:

• Herzog et al. (PLoSOne 2012;7:e29851) commented on the poor concordance of lipidomics measurements in seemingly similar samples and take the Oresic study as an example. *"Despite considerable efforts, the concordance between the lipidome composition and quantities of individual species obtained by different analytical approaches from similar samples remains poor."*

• Brezar et al. (Endocrine Rev 2011;32:623) describe in an authoritative review in great detail the main observations of the JEM study and speculate the mechanisms of diabetes pathogenesis they could reveal. The end their comment by stating: "*These observations await independent confirmation*."

• Ziegler and Nepom (Immunity 2010;32:468) consider in a review the study "*a preliminary report*".

• The review of Eisenbarth (Diabetes 2009;59:759) presents the findings and concludes: "*Both a power and potential weakness of these studies is that the analysis of multiple parameters, and thus testing of thousands of hypotheses at the same time, will yield many false-positive results. Thus, replication studies will be crucial*."
• Madsen et al. (Anal Chim Acta 2010;659:23) discussed the limitations of the study: "*A recent study by Oresic et al. [103] has shown that specific metabolic signatures are present before the onset of T1D, opening possibilities for treating patients before complete insulin dependence has developed. The study relied on only univariate testing (Wilcoxon) in order to find differential metabolites; leading to list of possible markers, but presenting no good discussion of their relationship to each other or to the disease. The study did not evaluate the diagnostic properties of the metabolic perturbations, but the results are none the less interesting and deserve further investigation".*
In conclusion, the findings were considered interesting but preliminary, and needing further confirmation. Their validity of the findings was not challenged in any of the citations.

Independent repetition of findings of the JEM study

It is also important to examine whether the findings of this study have been independently reproduced during the 8 years following its appearance, which would provide further credence for them. The finding that the metabolomics profile of diabetic patients is different before and after seroconversion has been reproduced by several studies [e.g. Pflueger et al. Diabetes 2011;60;2740 (Oresic as co-author); La Torre et al. Diabetes 2013;62:3951 (Oresic as co-author); Lappas et al. Dibetologia 2015: 58:1436; Araujo de Pina Cabral et al. Diabetol Metab Syndr 2015;7:52]. The metabolomics analyses of the La Torre et al. study appear to have been carried out in the lab of prof. Oresic, but using blinded samples in random order, so manipulation of the results is not possible. There is considerable variability between details of the individual findings, as is typical of metabolome studies with small size, but the phenomenon itself seems to be reproducible. Likewise the phenomenon of metabolomics changes before and after seroconversion has been reproduced in several studies on animal models, some their own (Sysi-Aho et al. PLoS Comput Biol 2011; Oct 7, e1002257), but also by others (e.g.; Åkesson et al. Metabolomics 2011;7:593; Madsen et al. PLoS One 2012;7:e35455; Grapov et al. Metabolomics 2015;11:425; Overgaard et al. Metabolism 2016;12:13).

Detailed analysis of the JEM paper

Just by reading the JEM article it is impossible to judge whether there is data tampering or fabrication somewhere along the study. This would need access to the laboratory log books, and careful scrutiny of all experiments carried, the raw data obtained and whether all of them were included in the final data set as well as reasons why some may not have been included. Examination of the published article only allows evaluation whether the data are presented, described and interpreted in scientifically acceptable fashion. This must be taken into account as a limitation of my evaluation. I go next through every section of the publication, mainly focusing on details that may deviate from good scientific practice.

Abstract
The statements made in the abstract are largely backed by the results, although some corners were cut in this brief presentation of the key finding. For example, only one type of phosphatidyl choline (PC) was significantly reduced at birth and some claimed differences in relation to seroconversion were not statistically significant. The final conclusions elucidating the potential importance and applicability of the results are acceptable.

Introduction
Introduction ends with: '*We…provide evidence that metabolomics disturbances precede the autoimmunity that is characteristically observed before development of type 1 diabetes'.* The statement could have been a bit more cautious (e.g. our findings suggest that…), but I do not see this an overt exaggeration.

Results

Subjects and Analytical Platforms: The way subjects and controls were selected (Fig. 1) from two studies and samples analysed in different numbers and at different time points, and in relation to seroconversion, is somewhat confusing but understandable after careful reading. The fact that different children contributed samples, from random subsets, at the different ages (from birth to 10 yrs) probably contributed to the large variation of the results.

Case report and Fig. 2: The public critique of the publication has focused on Fig. 2 concerning its representativeness and claims for overinterpretation of its meaning. It is a case report of longitudinal changes of selected metabolites and autoantibodies in a girl developing T1DM. The first sentence of the paragraph describing this data states that the figure '*demonstrated the interdependence of metabolic and immune system factors*'. Although it is nowhere stated that this is a representative case, it is implied by the text. I attempted to verify this by comparing the data in Fig. 2 with the compiled data presented elsewhere in the text. We can have a look at the 8 metabolites presented in the figure one by one (**bold text** when the levels in Fig. 2 are not in agreement with compiled data):

1. lyso PC:
• high before seroconversion in Fig. 2B
• high [PC (18/0:0)] between 0-2 yrs in Fig. 3
• high [PC (18/0:0)] before seroconversion in Fig. 5
2. ether PC:
• low **before** seroconversion in Fig. 2B
• low between 1-2 yrs in Fig. 3 + sporadically thereafter
• low mainly **after** seroconverson in Fig. 5
3. ketoleucine:
• low before seroconverision in Fig. 2B
• no difference in Fig. 4
• low before IAA seroconversion in Fig. 5C, no other differences
4. leucine:
• high peak before seroconversion in Fig. 2B, additional peaks around seroconversion
• variable non-significant levels in Fig. 4
• high level before and low after seroconversion in Fig. 5B, similar trend in Fig. 5C
5. glutamine:
• low **before** seroconversion in Fig. 2C
• no changes in fig. 4
• low **after** seroconversion in Fig. 5B (variable differences in 5C)
6. a-ketoglutarate:
• low **before** seroconversion in Fig. 2C
• low at birth, no significant differences thereafter in Fig. 4
• **no significant differences** before and after seroconversion in fig. 5
7. ketoleucine:
• low before seroconversion in Fig. 2C
• no differences in Fig. 4
• low before IAA seroconversion in Fig. 5C
8. glutamic acid:
• high before seroconversion in Fig. 2C
• no differences in Fig. 4
• high before seroconversion in Fig. 5C

Moreover, the text states changes in succinic acid, alanine and 2-hydroxybutyric acid, but the results are not shown. To summarise, the changes reported in Fig. 2 for 5 paramenters (lyso PC, ketoleucine, leucine, ketoleucine and glutamic acid) were supported by the complied data of these metabolites in other figures of the paper. Those of 3 (ether PC, glutamine and a-ketoglutarate) were not reproduced in the compiled data. Taking the high variability of the data it is quite expected that if we take an individual subject, the changes in all metabolites do not follow the pattern found in the compiled data. The data of Fig. 2 look therefore

quite real. It is unfortunately a common practice in scientific reporting that a 'typical case' is in fact the 'best case'. This is the likely explanation why this not completely convincing case was chosen to illustrate the metabolite vs. seroconversion pattern. The authors should have explained in more details the reasons for selecting this case, including its limitations. I would classify my critique of this Figure as superficial scientific reporting. It is tenuous to classify it purposefully misleading, because the text makes no direct claims on its significance. If I try to describe the figure, it shows an example how metabolic changes may progress in an individual before seroconversion, without excluding the possibility that similar changes can be observed, though at statistically lower frequency, in controls. An informed reader should be able to put the message of this figure into correct perspective.

My opinion about Fig. 2 is somewhat less judgmental than those of Tale Oy and 4Pharma Ltd who have considered the figure misleading because similar changes were seen in the two controls of this case. With one case and two controls very little can be said about their general representativeness for the whole study population. Only comparison with the complete data set will tell whether a single case is representative. IN my comparison 5 out of the 8 parameters shown in Fig. 2 followed the patterns seen in the compiled data. How many findings must be similar for the sample to be considered representative? As said before, this was probably the most representative case. Similar changes in two controls do not provide definitive proof of the contrary.

Fig. 3 appears to be largely correctly interpreted and described in the Results section. There is one incorrect statement: sphingomyelins are stated to be consistently decreased although the p-value was <0.09 (however, <0.05 in Fig. 3!).

Fig. 4 has 7 x 75 = 525 individual and independent measurements. Of them 26 are stated to differ statistically significantly between cases and controls. Incidentally, this is exactly 5% of the comparisons, which is the same as the likelihood of finding a difference by chance. There were 23 samples with undetected levels. Their exclusion from the calculations does not essentially change the outcome (5.18% positives). Therefore my conclusion about the data of Fig. 4 is that there are no genuinely significant differences between cases and controls there. The number of the positive findings is identical to the number of positive findings one would make by chance. This also seems to be the authors' conclusion: *'...no clear persistent changes were observed in small molecule profiles...'*. The significance of these findings does not seem to be overemphasized.

Fig. 5 compares the selected metabolite differences between cases and controls before and after seroconversion (+ 18 mo) irrespective of age. The statements in the text about changes and differences were largely supported by the data in the figure, although some variability was in the significance of changes when the data were stratified according to specific seroconversions. The claim of lysoPC changes is only partly supported by the data. One difference was stated as increase at $p > 0.05$ (glutamic acid).

The presentation of the supporting online material in Figs. S1-S9 is largely acceptable.

Discussion
Paragraph 1: *"Our study strongly suggests that metabolic dysregulation preceded overt autoimmunity in type 1 diabetes".*
It is a matter of taste whether the data 'strongly suggest' or just 'suggest' the dysregulation, but even the former wording does not qualify for scientific fraud or even unrealistic conclusion.

Paragraph 2: "*…the children who developed T1DM have reduced serum succinic acid and PC at birth and reduced levels of multiple phospholipids and triglycerides throughout followup*".
• succinic acid: based on unreliable data of Fig. 4 – may be a chance finding
• PC: slight overstatement, because only one of seven PC's was significantly reduced
• phospholipids and triglycerides: supported by data
"*diminished choline values in progressors at birth*"
• true if they mean phosphatidylcholine, no data on choline are in the results

Paragraph 3: *"these PCs were consistently low in the progressors"*
• There is a slight overstatement here. The levels were low but not consistently.

Paragraph 4: *"The appearance of each islet autoantibody was preceded by elevated concentration of lysoPCs"*
• correct

Paragraph 5: *" The appearance of autoantibodies against insulin and GADA…were preceded by diminished ketoleucine, elevated BCAAs and elevated glutamic acid"*
• ketoleucine: correct only for insulin
• BCAAs (leucine, isoleucine, valine): no statistically significant elevation
• glutamic acid: correct

Fig. 6 is plausible but highly speculative, because most of the changes described by the boxes are based on non-significant differences. Speculations are allowed in Discussion, but the most vindictive statement in figure legend is 'Detected changes' because most of the stated changes are not supported by statistical significance, which means that they were not 'detected changes'.

The described changes of the metabolites before and after seroconversion in GADA are as follows:

Metabolite	change in Fig. 5	claim in Fig. 6
Leucine	no significant change	from high to unchanged
Ketoleucine	no significant change	from low to unchanged
Alanine	no significant change	from unchanged to low
Succinic acid	no significant change	from low to unchanged
a-KG	no significant change	from low to unchanged
Reducing equivalents	not described (?)	from low to unchanged
Antioxidant lipids	not described (?)	from low to unchanged
Glutamic acid	from high to unchanged	from high to unchanged
Glutamine	from high to unchanged	from low to unchanged
GABA	no significant change	from high to unchanged

It thus seems that Fig. 6 is highly speculative and not supported by the data of Fig. 5. It unnecessarily reduces the integrity of the presentation, and it is surprising that the journal accepted it.

Paragraphs 6 and 7: They discuss here the therapeutic implications of the findings. I have nothing to criticize.

Paragraph 8: "*it is essential that these results are validated using other well characterized population cohorts*", which "*will hopefully, in time, be able to confirm or reject our findings*."
These statements moderate markedly the tone of the conclusions and put their current value into the right perspective, rather as a hypothesis generating set of data than an established fact.

Paragraph 9 states as a finding of the study that "*dysregulation of metabolism precedes beta cell autoimmunity in overt type 1 diabetes*." The evidence is a bit shaky, and I would have presented the conclusion more modestly and conditionally. Furthermore, they state that the findings "*imply that metabolic or immunomodulatory interventions during the preautoimmune period may be used as potential strategy for prevention of type 1 diabetes*". This is normal speculation presented in Discussion. I find it acceptable. It is normally the task of the referees and editor to moderate too excessive conclusions in Discussion.

Materials and methods
I have no other comments to make on this than the missing statement of customary blinding and randomization of samples, to eliminate concerns about operator-dependent bias in the results.

The authors may have ‘shot themselves in their own foot’ by doing too primitive statistical analysis of the data, by just comparing the metabolite levels vertically, using Wilcoxon rank sum test, to compare converters and controls in cross-sectional manner. A longitudinal comparison of changes during seroconversion could have revealed changes that now remain hidden. Another much more powerful statistical approach would have been principal component analysis. As stated by the statistical evaluations, the correction of significances for multiple comparisons and false detection rates was at least partly neglected. However, the Tale and 4Pharma evaluations considered the statistical analysis largely acceptable.

Summary

I did not find in the publication signs of data manipulation, neither gross miss- and overinterpretation of results or unrealistic conclusions. The former could be detected by scrutiny of the original laboratory log books and raw data. This comparison has been made by the evaluators of Tale Oy and 4Pharma, and no important irregularities in this respect were reported. Manipulated data are typically ‘too good to be true’. This is not the case with the current paper, the results are merely ‘too modest to be false’, and qualitatively quite expected. The study is by no means perfect. It is based on a very small material to be analysed for parameters with notoriously large interindividual variability, resulting, as expected, in marginally meaningful findings. The statistical analysis of the data was rather primitive with low power. More sophisticated statics could have revealed much more from this complex data set. I identified several erroneous interpretations in Results and Discussion, stating non-significant changes as changes (at $p<0.06$ and $p<0.09$). There is therefore some evidence of superficial data interpretation, because all conclusions did not strictly adhere to the results. I found the highly criticised Fig. 2 in the context of this publication acceptable, but its limitations were insufficiently explained. The discussion is in places rather speculative (e.g. Fig. 6), but mainly realistic, taking into account the preliminary nature and small size of the study. In my opinion the paper does not contain unrealistic conclusions. The main message of the study, i.e. that some metabolomics changes appear to occur in children before seroconversion, is not jeopardised by the flaws in details that I pointed out.

To conclude, I was not able to detect unequivocal proof of fraud (data tampering, falsification, or erroneous reporting) in this publication. Neither did it contain such excessive interpretation of results or unrealistic conclusions that would deviate from the normal writing style of scientific publications.

Sincerely,

Ilpo Huhtaniemi