Assessment of scientific misconduct

Ville-Petteri Mäkinen
April 1st 2016

Article under assessment

Oresic M, Simell S, Sysi-Aho M, Näntö-Salonen K, Seppänen-Laakso T, Parikka V, Katajamaa M, Hekkala A, Mattila I, Keskinen P, Yetukuri L, Reinikainen A, Lähde J, Suortti T, Hakalax J, Simell T, Hyöty H, Veijola R, Ilonen J, Lahesmaa R, Knip M, Simell O. Dysregulation of lipid and amino acid metabolism precedes islet autoimmunity in children who later progress to type 1 diabetes. (2008) J Exp Med 205:2975-2984.

Assessment procedure

The letter from VTT defined my task as "an informed and independent second opinion as to whether there is unequivocal proof of fraud (data tampering, falsification or erroneous reporting)" in an article written by Matej Oresic and co-authors in the Journal of Experimental Medicine (JEM) in 2008. To accomplish the task, I designed an assessment process that would cover the essential aspects that, in my opinion, would be relevant to the case.

First, I read the published research article as it appears in the journal and evaluated its scientific quality and possible red flags for misconduct (Chapter 1). I did not read any of the other material VTT provided me at this stage, or had any contact with the other evaluators, colleagues or the authors to ensure an impartial review.

In the second step, I compared my assessment of the scientific quality with those of the peer-reviewers, and the previous assessments that were commissioned by VTT. The goal of this step was to identify any areas of ambiguity, and to confirm that all scientifically relevant issues would be appropriately covered (Chapter 2). It also gave me a chance check if the article omitted critical details that were revealed by the re-analyses by independent statisticians.

Lastly, I used all the material provided by VTT to answer three questions: i) were any of the original data likely to be fabricated, ii) were any of the results likely to come from tampered analyses, and iii) were any of the conclusions derived from intentionally misrepresented findings (Chapter 3).

1 Review of scientific content and quality

1.1 Study design and methodology

The article "Dysregulation of lipid and amino acid metabolism precedes islet autoimmunity in children who later progress to type 1 diabetes" by Matej Oresic, Olli Simell and co-authors reports findings from a multi-year study of children who have a high genetic risk for type 1 diabetes. The study focused on 56 children who progressed to diabetes before the end of the follow-up period, and 73 children who had similar genetics and characteristics, but did not develop diabetes. Almost all children were a subset from a much larger cohort (Type 1 Diabetes Prediction and Prevention Study initially screened >100,000 newborns). In the article, the authors report the results from detailed analyses of circulating lipid molecules in blood from multiple time-points.

The onset of a complex yet rare disease such as type 1 diabetes is difficult to study in humans due to the sporadic occurrence and several years of non-symptomatic incubation period. When type 1 diabetes is diagnosed, the condition causes secondary symptoms that mask any original causes, so any early clues on the disease mechanisms are vital to develop preventative strategies. To this backdrop, the rationale and study design are appropriate as the paper focuses on individuals with high genetic risk to ensure higher proportion of participants who will get diabetes (cost-effective approach for rare diseases), and the molecular data is obtained *before* the disease onset to investigate the non-symptomatic early phase of the disease process.

Most of the data were obtained from serial blood samples that were collected according to standardized procedures at pre-defined intervals. The authors confirmed that storage time before analysis did not affect the results, and therefore the presented data is unlikely to represent technical sample handling artefacts. Although I am not an expert on the specific technical details of mass spectrometry (the method of measuring lipids in blood), based on other similar articles in the literature that I have reviewed in the past, the analytical procedures seem reasonable and appropriate safeguards were in place to ensure high quality data.

Statistical analyses were conducted for each lipid measure separately or by aggregating the concentrations of related lipid molecules into a few classes, and then testing the classes separately. Average values of the original or aggregated measures were then compared between the children who developed diabetes, and those who remained healthy, across all available time points to detect the earliest divergence that would predict the disease onset. The authors used established algorithms (univariate tests and linear regression models) to evaluate whether a divergence could be explained by chance or whether it was likely to arise from a "true" difference. The statistical methods were appropriate for the study. The bulk of the results were depicted by coloured heatmaps (Figures 3 and 4), and any pixels that satisfied a pre-defined statistical threshold for a difference between progressors and controls were highlighted. Figures 2-4 were legible, but they were also packed with too much information, which made it difficult to interpret the take-home message.

1.2 Comments and criticism

The authors omitted all numerical results on the statistical findings from the Abstract. Therefore, it was very difficult to assess the robustness of their claims on the lipid findings. The numbers were presented in the Result section in the main text, but in my opinion it is not good practice to omit these critical indicators from the abstract.

The Introduction is well written and highlights the important reason for conducting the study: the events that lead to type 1 diabetes are poorly understood and the discovery of effective prevention remains a challenge.

The recruitment flow chart and study design is depicted in Figure 1. While the chart itself is clear, the specifics are not described in the text, nor are they covered in the Discussion. For instance, it is unclear how many of the 50 progressors from DIPP were also participants in the nasal insulin trial. It is possible that the insulin treatment could affect parts of the lipid profile, and confound the results.

The analytical platforms produced wildly different numbers of lipid measures, which is somewhat surprising, as one would not expect to see such dramatic differences in the distribution and concentrations of lipid species in serum (RESULTS: Analytical platforms). However, it is still plausible that such differences can arise due to detection limits and inherent properties of the source materials (cord vs. venous blood).

The statistical results regarding Figure 3 are poorly described. It is not immediately clear what type of analysis was used to produce the P-values. I assume it is the Wilcoxon rank-sum test for the lipid classes similar to the heatmaps in the figures.

Overall, the P-values could have been reported better. In colloquial terms, a P-value is a measure of how certain we can be that the observation is true instead of a random chance. If $P < 0.05$, we accept that there is a <5% chance that the difference is a random fluctuation. In metabolomics studies such as this, the large number of independent observations (i.e. multiple tests) complicates things: if you perform a large number of tests, then it is likely that at least a few of them will produce a P-value below 5% purely by chance even when the data are completely random. In the method section, the authors provide technical details on false discovery rates (FDR), which is a method to adjust the statistics against multiple testing. The authors do not explicitly state it, but I assume they used the Benjamini-Hochberg procedure to calculate FDR which, in my experience, tends to be overly optimistic. Specifically, FDR estimates in these types of studies tend to give unreliable and artificially low false discovery rates for the top signals, and I personally prefer using the conservative Bonferroni adjustment as the threshold for a robust finding along with a heuristic rule of thumb ($P < 0.001$) for a statistically plausible signal in a typical metabolomics study.

The main text only reports single test P-values without appropriate adjustments. Given that the P-values were not adjusted, the results may have to be assessed more conservatively. I would not put much weight on any findings with a single test $P > 0.001$, which means that a large proportion of the result section should be considered suggestive rather than statistically significant.

In the Discussion, the authors provide a comprehensive interpretation of their results and develop hypothetical scenarios based on existing knowledge of biochemical and cellular processes to explain the observed lipid and metabolite associations with diabetes progression. This is a standard feature in metabolomics studies, and all the authors' conclusions and proposed explanations are reasonable if you accept the statistical findings. Even if you do not accept that the findings are sufficiently supported by the statistics, the conclusions can be passed as informed speculation on the suggestive trends that were observed in the study. Importantly, the authors acknowledged that further validation will be required to confirm the observations.

1.3 Controversy surrounding Figure 2

Before I accepted the assignment, I found out about the controversy concerning the case study that was illustrated in Figure 2, and then further developed into a hypothesis in Figure 6. Within the context of the paper, I regarded Figure 2 as a best-case-scenario where a single individual shows something very clear before the autoantibodies emerge, followed by a conspicuous restoration of a more stable profile after the autoantibody peaks. Such anecdotal evidence is common in medicine. Admittedly, it fits with the authors' story extremely well, and can be exploited to create hype around the methodology, but after reading the full text, it is easy to see, in my opinion, that the effects in the other individuals were nowhere near as straightforward. As a literary work, the paper has the structure of a tabloid: the headline and abstract make exciting claims that are diluted by the actual contents. Speculation and generation of hypotheses is part of science: it is possible that the observation could be true, but unique to the particular girl, which would explain the disconnect with the other participants – this possibility is the reason why anecdotal evidence and case studies are reported in the first place even though they cannot be substantiated by statistics.

2 Comparison across multiple reviews and re-analyses

2.1 Peer-review

The process from the first submission to a journal and the acceptance for publication in JEM took close to two years. Five top-tier journal rejected the manuscript (they have very high thresholds for publication), and this probably explains the prolonged process. I was provided with two reviewer reports from the JEM submission.

Reviewer 1 expressed concern about the size of the study, and also wondered if Figure 2 was really representative of the full datasets. I share these concerns, although I thought it was evident from the group statistics that the case study was not a typical example of the data. The reviewer was also questioning the clinical relevance of the study, again a consequence of the relatively limited numbers and lack of replication in an independent dataset or experiment. The biological explanations that the authors provided in the Discussion were criticised by the reviewer (which I missed, probably due to lack of expertise in cell biology). The reviewer concluded that while intriguing, the findings would need to be investigated further to make them convincing, an opinion that I share. Many of the specific comments by Reviewer 1 are minor, but there were recurring themes. In particular, the reviewer, like I, criticized the lack of multiple testing adjustments, and questioned the statistical robustness of the separation between diabetes progressors and healthy controls.

Reviewer 2 was excited about the study, and thought it was ground-breaking. He or she was impressed by the case study, and was willing to accept the manuscript without mechanistic validation. While I also think these type of papers are worthy of publication without further experiments, I do not think that the results here were as impressive as the reviewer suggested. Based on my experience both as an author and a peer-reviewer, it appears to me that Reviewer 2 may not have read or investigated the article in sufficient detail, and based the comments on the title, abstract and figures rather than the full text. Unfounded hype is the undesirable consequence of promising too much in the title and abstract, as casual readers may not have the time or expertise to carefully examine the contents, especially when the study design and other figures were fairly complicated, and the field of study was still in its infancy.

2.2 Re-analyses by independent statisticians

VTT commissioned two independent experts to re-analyse the lipidomics and metabolomics data in 2014. The two statisticians re-analysed the lipidomics data with the same methods used in the article, and both confirmed the authors' calculations as reasonably accurate. Not surprisingly, both experts criticized the lack of proper adjustments for multiple testing. They nevertheless found that the data analyses in the article were technically correct for the most part and the strongest signals could be reproduced. As expected, Figure 4 was deemed unreliable: I counted 25 out of $7 \times 75 = 525$ pixels that had a significant P-value below 5%. This is very close to the expected number of false discoveries so it is impossible know if any of the signals were real.

Both experts concluded that the case study in Figure 2 had some abnormal characteristics, but that the girl’s data did not represent a discrimination between progressors and controls or any consistent temporal pattern of type 1 diabetes development. In particular, most of the lipid and metabolite concentrations from her samples were similar to the control population. This was impossible for me or anyone else to verify based on the article alone. I agree that there is no statistical support for any claim that the girl represents a typical case of type 1 diabetes onset (in the paper, she was labelled as an example with a long follow-up). That said, anecdotal evidence or peculiar outliers can lead to new discoveries, so I would not dismiss Figure 2 simply because it depicts a single individual. It is conceivable that the authors could have avoided the controversy by presenting specific analyses on how different the overall metabolic profile of the girl was from the rest of the study cohort.

There was one important piece of information that should have been included in the text. Neither I nor the two peer-reviewers realized the incompleteness of the time series data. After reading the two statisticians' reports, it became clear that the lack of multiple serial blood samples for many of the

children, and the patchy follow-up periods, had a detrimental effect on the statistical modelling. I can now see the small numbers beneath the heatmaps in Figures 3 and 4, but during the review they were lost in the complexity of the plots since they were not explicitly mentioned in the text. It is difficult to say if disclosing this weakness in the study design in the Discussion would have prevented the publication in JEM, but it is, in my opinion, problematic not to adequately describe this type of limitation. If the weakness would have been explicitly dealt with, I personally could have still accepted the study for publication.

3 Conclusions

3.1 Did the authors fabricate data?

I found no suspicious features in the article that would suggest any of the samples, measurements or molecular concentrations were fabricated. All the sample procedures and technical details are consistent with accepted practices in the field, there are no outrageous claims of superior accuracy and the lipid species included in the study were, if anything, conservatively filtered to include the best quality portion of the data. My overall impression from the hundreds of concentrations curves provided by VTT fits with the nature of typical metabolomics data that come from clinical samples. The age and seasonal distribution of the participating children are unremarkable, although based on the limited information provided it is impossible to say if they represented typical children in the population-based parent cohort.

3.2 Did the authors tamper with statistical analyses?

I found no convincing evidence that any of the statistical analyses were falsified so as to promote an erroneous conclusion. There were several technical issues that need attention, but it is important to remember that the first author (as far as I know) is a metabolomics expert, not a statistician, and the majority of the other authors are clinicians or basic biologists. Although I have not followed the case in the media in detail, nor do I have first-hand knowledge of what happened at VTT, the background information I do have points to communication problems and the subsequent disintegration of the research group. To this backdrop, it is perhaps not surprising that the quality of the work may have suffered in the toxic environment, which would easily explain the technical problems in the article. The partial re-analyses by the two independent statisticians corroborated the strongest findings, which is further indication against deliberate tampering.

3.3 Did the authors deceive readers by misrepresenting data?

I found evidence of poorly represented results in the main text, and exaggerated claims in the title and abstract, but I cannot unequivocally state that they constitute a fraud. There is a reasonable possibility that some of the problems were due to lack of experience and some were due to the break-down in cross-disciplinary communication. It is important to remember that human metabolomics studies were very much an emerging field in 2008, and there was no collective effort to improve the reliability of articles. Indeed, I still get manuscripts for peer-review with grandiose claims in the title only to find that the statistical analyses are poorly conducted and the interpretations have no logical basis. In most cases the peer-review system works and authors improve the quality of their work, as happened here with the first five rejections by the top-tier journals. While some readers might get the wrong impression from a superficial reading of the article by Oresic and co-authors, I am of the opinion that despite the bombastic first page and technical issues, there is enough information within the text and figures to get a balanced view of the findings.