EVALUATION REPORT

Study: independent external evaluation of article

Dysregulation of lipid and amino acid metabolism precedes islet autoimmunity in children who later progress to type 1 diabetes

with regard to the correctness of its data analyses and conclusions.

Version	Date	Changes
Ready	18 March 2014	

Confidential

This document is the property of Tale Oy and the customer company (VTT) and may not be divulged to external parties fully or in parts without a written authorisation from the above-mentioned party.

CONTENTS

DEFINITIONS

Article, Publication	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. J Exp Med. 2008 Dec 22;205(13):2975-84. doi: 10.1084/jem.20081800
Study	The "original" scientific study, including methodology, results and conclusions presented in the article
Study plan	
Evaluation	Evaluation and analysis of the correctness of the data analyses and conclusions in the study, carried out by an independent, external party.
Evaluator	Person conducting the evaluation.

1 FOREWORD AND OBJECTIVES

Commissioned by VTT, Tale Oy/Tommi Nurminen has carried out an independent evaluation of the reliability and correctness of the data analyses and conclusions in the article. The evaluation was based on the Finnish Advisory Board on Research Integrity's guideline "Responsible conduct of research and procedures for handling allegations of misconduct in Finland".

Assistance was received in statistical programming from Risto Heikkinen of Statisti Oy.

The evaluator does not have a personal relationship to the authors of the study and no prior part in the study in question. It should be noted that the evaluation was carried out by a person with education and work experience in statistics. The evaluation focuses on the correctness of data processing and statistical analyses. The evaluation of the final conclusions and the cause-effect relationships may remain superficial, as this would require more in-depth medical expertise.

1.1 VTT's questions to the evaluator

Question 1: Sufficiency of the time series

The abstract of the article states the following: "Serum metabolite profiles were compared between sample series drawn from 56 children who progressed to type 1 diabetes and 73 controls who remained nondiabetic and permanently autoantibody negative." How many persons have sufficiently long time series for metabolites (metabolomics) and lipids (lipidomics) in order to be able to draw reliable conclusions?

Question 2: The selected case's representativeness and difference compared to the controls

Figure 2 presents metabolomics and lipidomics results for a girl who developed diabetes. A certain application states the following concerning this figure: "A representative example of metabolomics in the DIPP study, showing selected autoantibody and metabolite changes during the prediabetic period in a girl who progressed to T1D at close to 9 years of age." Is this figure a representative example of the metabolomics results?

Figure 2 does not present any controls. Could the curves presented in Figure 2 be repeated for all patients in the data and the selected controls in the data? Is there significant difference in the trends of the patients and the controls?

Question 3: Reliability of the results as a basis for a predictive model

Are the results clear enough to be used for predicting the onset of diabetes on an individual level? Is the presented data sufficient for generalising the peaking of the presented markers before the seroconversion of the antibodies? Are the conclusions in Figure 6 in line with the results in the article?

2 RESEARCH PLAN AND DATA

2.1 Description of the research based on the article

The article describes a research studying whether serum lipids or other metabolites can be used to predict type 1 diabetes (T1D) at an early stage. The research setup is a retrospective case-control study.

The data comprises follow-up data gathered from three Finnish university hospitals (Type 1 Diabetes prediction and prevention study [DIPP]). Children belonging to the genetic risk group participated in regular follow-up tests that included blood samples. 50 cases, i.e. children who had contracted T1D, ended up in the study. One or two controls were chosen for each case from the T1D risk group of the DIPP register – children who tested negative to autoantibodies and did not contract T1D during the study. The controls were matched according to time and place of birth, gender and the HLA risk group. There were a total of 67 controls in the data.

The study also included data from six children who were part of the follow-up of the Turku-based STRIP study and had been diagnosed with T1D, and from their six controls matched by age and gender. Compared to the DIPP register, children in the STRIP register do not necessarily have the genetic risk factors for T1D. Such risk factors were identified in three of the six cases, however.

Blood samples from the cases and controls were analysed in four batches. Of a total of 1196 analysed blood samples, 53 lipids that could be measured from all blood samples ended up in the statistical analysis. Other metabolites were analysed from 13 cases and 26 matched controls. 75 metabolites ended up in the statistical analysis.

The study also involved the collection and analysis of placental blood samples. The related results are not a central part of the article, and the data is unavailable.The analyses in question are consequently excluded from this evaluation.

2.2 Description of the data and various observations concerning the data

The evaluation was carried out based on the final, "cleaned" observation data. The observation data contained the background variables and the autoantibody concentrations. The times and places of birth and the HLA risk groups on which matching was based were not included in the data. The data_lipidomics dataset contains values for 53 lipids and the data_metabolomics dataset contains values for 75 metabolites.

2.2.1 Number of research subjects

It is stated on page 2976 of the article that the comparison of metabolism profiles is based on the 50 cases in the DIPP cohort and their 67 controls. Later in the same chapter it is stated that metabolites were also compared between the six cases in the STRIP cohort and their controls. In total, therefore, the study has used data from 56 cases and 73 controls. The available lipid data does indeed contain data from a total of 56 cases and 73 controls.

The numbers of research subjects stated in Table 1 of the article – 55 and 73 – do not fully match the above information.

2.2.2 Follow-up period lengths

Table 1 of the article reports key figures for the age of diagnosis and the age when the seroconversion was detected. We also calculated the lengths of the time series.

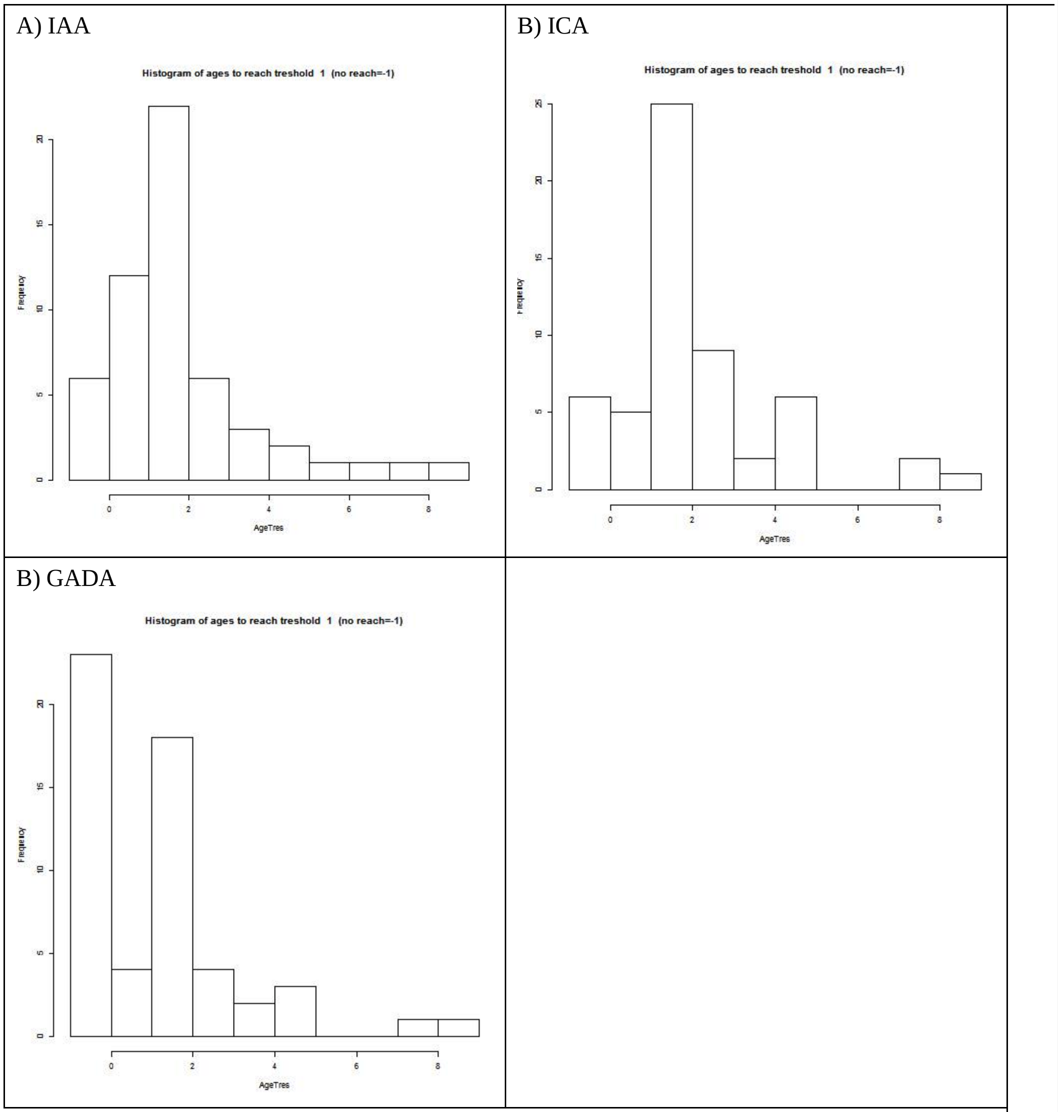
Table 1: Lengths of the time series

Time series		Cases	Controls
start from birth (day 0)	n	6	8
starting age (days)	median average	99.5 315	102 336
total length (days)	median average	1100 1299	1765 1808
time from first sample to the sample preceding seroconversion	median average	463 542	

Table 1 shows that the starting points of the time series were congruent between the cases and the controls, but the time series of the controls were longer, which is caused by the time series of the cases typically ending at the age of diagnosis, while the time series of the controls continued past the time at which the corresponding case was diagnosed. There are some exceptions: For case 41, the entire time series was from the time when the diagnosis was made. Case 49 also had numerous observations post-diagnosis. On the other hand, case 40 has no observations after seroconversion or diagnosis.

Figure 1 presents the age distribution of the occurrence of seroconversion and the GADA, IAA and ICA autoantibodies for the cases. For a majority of the cases, the autoantibodies were detected before they reached two years of age. Table 1 of the article states that the median age for the first detected seroconversion was 12 months.

Figure 1: Age at which the autoantibodies were detected


A) IAA

B) ICA

B) GADA

Children for whom the autoantibody level in question or a seroconversion was not detected are marked with -1

2.3 Individual notes concerning the observation data

Case 26 has only one observation at the age of 267 days. At that time, IAA in particular was elevated. It is not known when the subject was diagnosed.

Control -85 has a row in the observation data that suggests observations prior to seroconversion (age 2447 days). This is in conflict with page 2976 of the article, where it is stated that ”*nonprogressors who remained healthy and autoantibody negative*”.

Not all controls have a matched pair. Lipid data controls -25, -27, -32 and -94 are such controls. A matched pair is found for control -41 in the lipid data but not in the metabolomics data. The numbers of observations also vary in other ways between the datasets. For example, there are four observations for case 47 in the metabolomics data (until the age of 2350 days), but in the lipid data, the follow-up comprises ten observations, ending at an age of 4323 days (428 days before the diagnosis).

The article states that in the comparisons in Figure 3A of the article, the post-diagnosis measurements have not been included in statistical comparisons. The sample sizes reported in the figure, however, suggest that all observations have been included.

It would be unrealistic to expect that the article would describe the samples "missing" for various technical reasons in a very detailed manner. Considering the suspicions of data manipulation, it would be good to investigate the above-mentioned irregularities in the data, particularly the cases missing from the lipid data and the different lengths of the time series between the lipid and metabolism datasets.

3 GENERAL OBSERVATIONS ON THE STATISTICAL METHODS USED

3.1 Comparisons by age group and over the entire time series

The article describes very complex results based on a rather small set of data. The results are based on the "cross-sectional" comparison of the cases and controls, i.e., at each point of time, instead of attempting to evaluate the magnitude and predictability of the internal changes of the subjects. For example, when the increase in certain metabolite values before the occurrence of autoantibodies was evaluated, the comparison was done between the cases and the controls, instead of examining the increase of the values of the metabolites in question relative to the previous values for the same person.

The statistical method used in the per-age-group comparisons was the Wilcoxon rank sum test (or a Mann-Whitney U test). This test is a non-parametric test that is not sensitive to the distribution shapes. On the other hand, the method in question does not utilise other time points in any way whatsoever (not to mention other metabolites) in the estimation of the average value and variance. The result for each metabolite and time point is based solely on the data from the time point in question, and many differences caused by chance can be interpreted to be statistically significant. On the other hand, the article states an estimate of the relative share of false discoveries (*false discovery rate, FDR).* To be more precise, the article states the maximum q value, i.e., the maximum allowed share of false discoveries at which the observed differences (P<0.05) can be considered to be statistically significant. The observed significance levels are reported without consideration to other simultaneous tests and by simply reporting the estimated maximum q value.

Indeed, the comparison of some lipids has been carried out over the entire follow-up period using a linear mixed model, where the lipid values of the cases and controls are assumed to remain constant throughout the entire follow-up period. The internal correlation of individuals is modelled using a random effects model (resulting in a *compound symmetry* assumption). In this analysis, the results from one time point have a strong effect on the results from the other time points; in fact, the results are assumed to be the same and in this sense, there is the danger of circular reasoning with regard to the uniformity of the differences relative to time.

The differences between the cases and the controls are reported as ratios of the groups' median values by age group. This choice is somewhat odd, and in some cases it could be more sensible to compare, for example, the ratios of the geometric means.

3.2 Comparisons prior to the occurrence of autoantibodies

According to my understanding, the main result of the article is that metabolism changes are detectable in children who will contract T1D up to 18 months prior to the occurrence of autoantibodies. Although this result would certainly be material for understanding the biological mechanisms, it is likely to be less material with regard to predicting the occurrence of autoantibodies: The correct timing of the diagnostic test would require knowledge of when the autoantibodies would occur if they occur at all, which is naturally impossible. The predictive model must therefore be based on the metabolism differences between the cases and the controls that remain relatively constant at least until the occurrence of the autoantibodies or, alternatively, the predictive model must be very precisely delimited (for example, metabolism values during the first year of life can be used to predict the risk of autoantibodies occurring at the age of three).

As stated in Section 3.1, the article's main focus is on results that are based on per-age-group comparisons instead of attempting to utilise the unknown "time before the occurrence of autoantibodies".

3.3 Use of controls

Due to the small amount of data, the results are sensitive to the selection of cases and controls and to the comparability of their lipid and metabolite data. (This evaluation cannot comment on the reliability and comparability of the concentration values derived from serum samples.) In principle, the case-control set-up using matched pairs ensures comparability. The statistical comparisons of the article do not utilise the matching information – i.e. do not use statistical methods that would take the pair-internal correlation/dependence into consideration. In theory, this may have the following effects:

- Power: The statistical power is decreased, as the comparisons take the overall variation into consideration instead of the pair-internal variation.
- Bias: The imbalance in the observation plan (that two controls had been selected for some cases while only one control had been selected for others, and that the follow-up times are of different length for the cases and the controls) may cause bias in the comparison. The comparisons have used control samples even from age points from which samples for the corresponding case are no longer available, because post-diagnosis data is unavailable or because measurements for the age range were unavailable for some persons.

The above-mentioned consequences are significant only if the matching has real meaning, i.e., the expected values of metabolites within the pairs are similar, while there are significant differences in the levels between pairs. As a layman, I find it unlikely that time and place of birth would have any significant effect on the metabolism values, particularly as the measurements have not been made at precise times during the different ages. The balance of genders and HLA risk groups in the data can be evaluated, if they are believed to have a material connection to the expected values of metabolite levels.

The imbalance of the observation plan may be the reason why the non-parametric comparisons have been carried out as rank sum tests instead of as a signed rank test that takes the dependency into consideration (*Wilcoxon signed rank test*). The dependency could be better taken into consideration by using, for example, parametric analysis with random effects (e.g. a linear mixed model).

3.4 The representativeness of case 40

From the perspective of the analysis, the most challenging part is to evaluate whether the pattern of the metabolites prior to the occurrence of autoantibodies of case 40 presented in the article represents some typical pattern. For example, lysoPC is reported to be elevated at the group level 18 months prior to the appearance of the first autoantibody. At the group level, the difference appears to exist only between the first and third years of age (article Figure 3). The median age for the appearance of the first autoantibody is 12 months (article Table 1), so in fact, "18 months prior" means as a rule "since birth". For case 40, the values are also at their highest immediately after birth. The article remains unclear with regard to whether this is a case of an increase in lysoPC prior to the occurrence of the autoantibodies or whether it is a difference that is observable already at birth and does not disappear with time. Certain values of case 40 are normalised only after autoantibodies have appeared.

Although modelling the metabolite patterns before and after the occurrence of the autoantibodies could be attempted using a very specific statistical model, the article has (understandably) settled for comparing the cases and controls as groups, one parameter and time point at a time. The observed differences between the groups have been interpreted as "typical patterns", possibly drawing on the metabolism profiles of case 40.

4 DETAILED STATISTICAL ANALYSES

4.1 Main questions

Based on the previous chapter, the following questions concerning the correctness of the statistical analysis can be formulated:

1. If the statistical group-level analyses of the data are repeated in the manner described in the article, are the results the same?
2. Based on a visual evaluation, can it be concluded that the metabolism profiles of the cases are of a similar type to that of a case subject (and different from the controls)?

Comment [T1]: Lauseen tarkoitus oli ”kun (minä tai muu) arvioija tarkastelee profiileja visuaalisesti, voidaanko todeta..” eikä arvioida visuaalisen arvioinnin luotettavuutta menetelmänä.

The replication of the analyses aims to verify that the analysis results have not been knowingly falsified, but that the described methods give the presented results. Sensitivity analyses aim to evaluate whether the results (and the conclusions drawn on the basis of them) are sensitive to the selected methods. If the sensitivity analyses give wildly varying results, then the results are scientifically questionable, and the possibility of a purposeful selection of methods cannot be eliminated.

4.2 Claims with evidence that can be evaluated

Because the three questions presented by VTT are at different levels of abstraction, they cannot be answered without defining concrete claims and evaluating their correctness based on the data presented in the articles and the observation data. See below for a list of the essential statistical and case-specific results of the article:

1. General: ”Metabolic dysregulation precedes overt autoimmunity in type 1 diabetes” [p. 2980, 1st column].

2. "The representativeness of the selected case (subject 40)": Although the article does not directly claim this to be the case, the questions provided by VTT mention that the metabolism profiles of case 40 have been referred to as typical/representative data for a child who will contract T1D. The claim being evaluated is whether the metabolism profile of the girl in question was of the same type as those of the other children who contracted T1D and, correspondingly, different to the profiles of the controls.

3. ’Low lipids’: Throughout the follow-up, the lipid values of the cases (phospholipids and triglycerices) were lower compared to the controls. [p. 2977, 2nd column and Figure 3C; also p. 2980, 2nd column]:

   ”*Compared with children who remained autoantibody negative, the serum lipidome of progressors showed a consistent decrease in triglycerides (P = 0.005) and multiple phospholipids, including ether PCs (P < 0.001), PCs (P = 0.04), and phingomyelins (P = 0.09), in samples obtained from infancy to early school years (Fig. 3, A and B). The difference in lipidome was observed before the autoantibodies appeared and it persisted throughout the entire age range covered.*”

   Furthermore, similar results of differences between groups are listed beginning on p. 2978 (2nd column, ”*One subclass of phospholipids*”), although these results apply to the first years of life instead of the entire follow-up period.

4. Metabolic changes before the autoantibodies and a diagnosis [p. 2979, paragraph ”*Metabolic changes…*” and Figure 5; also p. 2981, 1st column, 2nd and 3rd paragraphs].

   “To determine whether the metabolome abnormalities showed association with the emergence of slet autoantibodies, the metabolomes of the progressors and of the nonprogressors were compared using samples obtained from ± 18 mo from the emergence of the first autoantibody (Fig. 5 A). The lysoPC PC(18:0/0:0) was increased 1.5-fold within 9 – 18 mo before seroconversion (P = 0.03) and 1.3-fold within the succeeding 9 mo (P = 0.005; Fig. 5 B).

   Metabolomes in the progressors and nonprogressors showed several unexpected differences around the time when autoantibodies appeared. Glutamic acid was 5.2-fold and 32-fold increased 9 – 18 mo (P = 0.02) and 0 – 9 mo before seroconversion (P = 0.02), respectively. Branched chain amino acids (BCAAs) such as leucine and isoleucine were also increased before seroconversion, whereas ketoleucine concentration was diminished.”

5. The key results are summarised in the article's abstract, which aims to comment on each result.

4.3 Analyses and results

4.3.1 Differences in lipid concentrations

<u>Replication</u>

Lipid concentrations were compared between the groups by repeating the analyses presented in Figure 3A of the article (with the exclusion of "birth", because the data was unavailable). That is, we derived the age at which the sample was taken in years and selected the observation that was closest to the midpoint of each age group. According to the article, only pre-diagnosis measurements were taken into consideration for the cases, while all observations were taken into consideration for the controls. The comparisons were carried out with the Mann-Whitney U test. However, during replication it became evident that the number of compared cases would be lower than reported in Figure 3A if the post-diagnosis observations were excluded. The sample sizes could best be made to match each other if all persons meeting the age criteria were included in the comparison regardless of diagnosis (Figure 2).

Figure 2: Lipid comparisons between cases and controls by age group, taking all observations into consideration

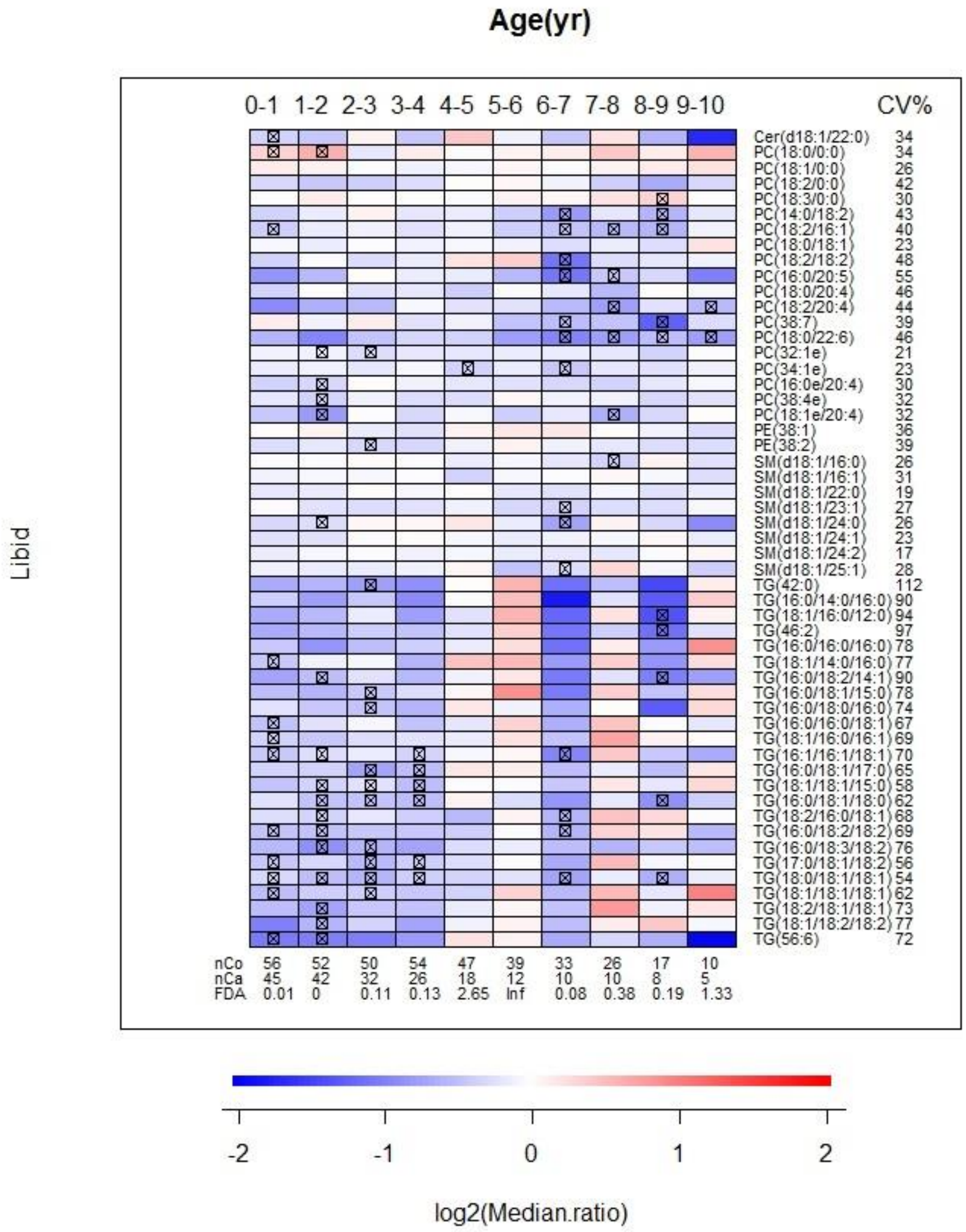


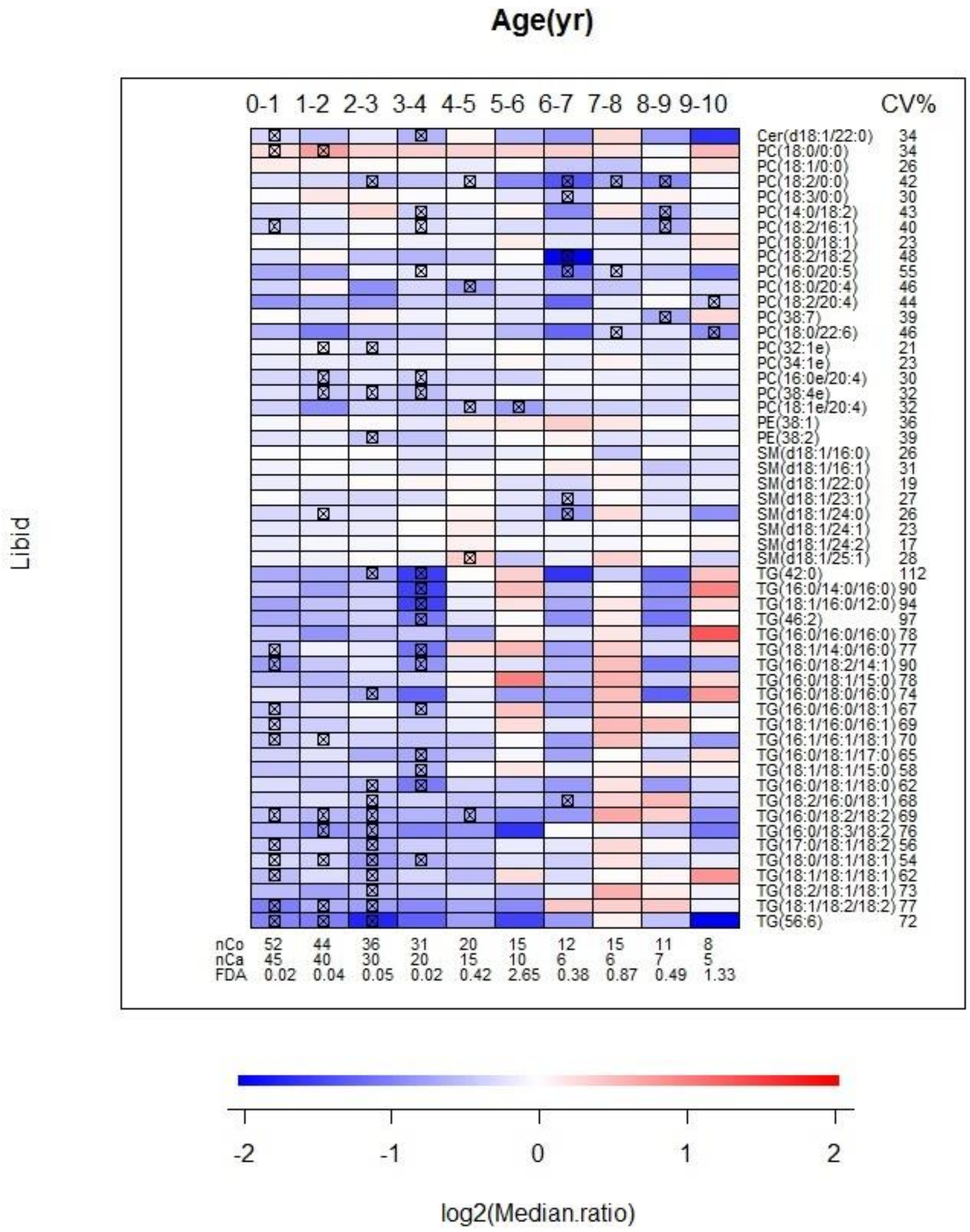
Figure 2 can be found to be identical to Figure 3A in the article. According to the Figure, triglycerides of the cases are typically at a lower level than those of the controls during the first years of life. There appear to be no differences later, or at least they are not as clear. The lysoPC value appears to be elevated during the first two years of life. From six to ten years of age, many PC values of the cases were low relative to the controls, although there is only a little data for these age groups, and based on FDR many of the observed differences are likely to be explained by chance. The early-age differences in triglycerides are more reliable.

Because the time series of the controls are longer in the data, the comparison of Figure 2 includes controls whose matching cases are not included in the comparison. For this reason, we repeated the corresponding comparisons with the post-diagnosis observations excluded. Observations where the age exceeds the matching case's age of diagnosis were excluded from the controls. This approach focuses the comparison on an interesting time period, that is, the time before the cases were diagnosed.

Compared to the above, the results do not change significantly; in the case of triglycerides, the differences even appear to strengthen. Further low significance levels ($P<0.05$) appear for triglycerides and PC values in the comparison for age group 3-4.

To summarise the replication of the analyses of Figure 3A of the article, it can be stated that repeating the described analyses using the available data achieves essentially the same results as described in the article.

Figure 3: Lipid comparisons between cases and controls by age group excluding the time after the age of diagnosis of the cases (and the matching controls)



Sensitivity analysis

An analysis corresponding to Figure 3B in the article was performed using linear mixed models of the measurements repeated by age group for data transformed to a logarithmic scale. The observations were selected as above, but the controls with no matching case were excluded from the analyses. The mean value of triglyceride concentrations was assumed to depend on age and whether it was a case or a control. Furthermore, we examined whether the differences between age and controls were dependent on age (interaction). The model took both the control pair and the individual person into consideration using random effects parameters. A Gaussian spatial covariance matrix was assumed between the observations: the stronger the correlation in observations, the smaller their temporal distance.

The results from the mixed models for triglycerides and etherPC support the result in Figure 3B of the article: there is a statistically significant difference in level between the cases and controls throughout the age groups (P < 0.001, cases at a lower level). The differences between the cases and the controls are not dependent on age to a statistically significant degree, which means that the comparison is based on the entire time series, as was also done in the article. With regard to etherPC, one case-control pair was excluded; its case had an anomalously large value at approximately 11 years of age.

The triglyceride values are at a lower level during the first year of age than later. EtherPC values appear to increase relatively evenly from the first year of age onwards.

To summarise, it can be stated that our results were in line with and led to the same conclusion as that presented in Figure 3B of the article.

Table 2. Triglyceride comparison (logarithm-transformed) using a linear mixed model

```
Fixed effects: log(Trigly) ~ Case + AgeG

                Value  Std.Error   DF   t-value p-value
(Intercept)  6.499102 0.06295946 1045 103.22678  0.0000
Case1       -0.185457 0.04811223   68  -3.85467  0.0003
AgeG2       -0.178567 0.05641959 1045  -3.16498  0.0016
AgeG3       -0.200233 0.06313278 1045  -3.17161  0.0016
AgeG4       -0.184310 0.06704615 1045  -2.74900  0.0061
AgeG5       -0.323104 0.07288576 1045  -4.43302  0.0000
AgeG6       -0.327986 0.08029509 1045  -4.08476  0.0000
AgeG7       -0.264840 0.08721602 1045  -3.03659  0.0025
AgeG8       -0.209568 0.09525103 1045  -2.20017  0.0280
AgeG9       -0.077963 0.10765226 1045  -0.72421  0.4691
AgeG10      -0.162987 0.11611774 1045  -1.40364  0.1607
```

Figure 4. Triglyceride profiles for cases and controls

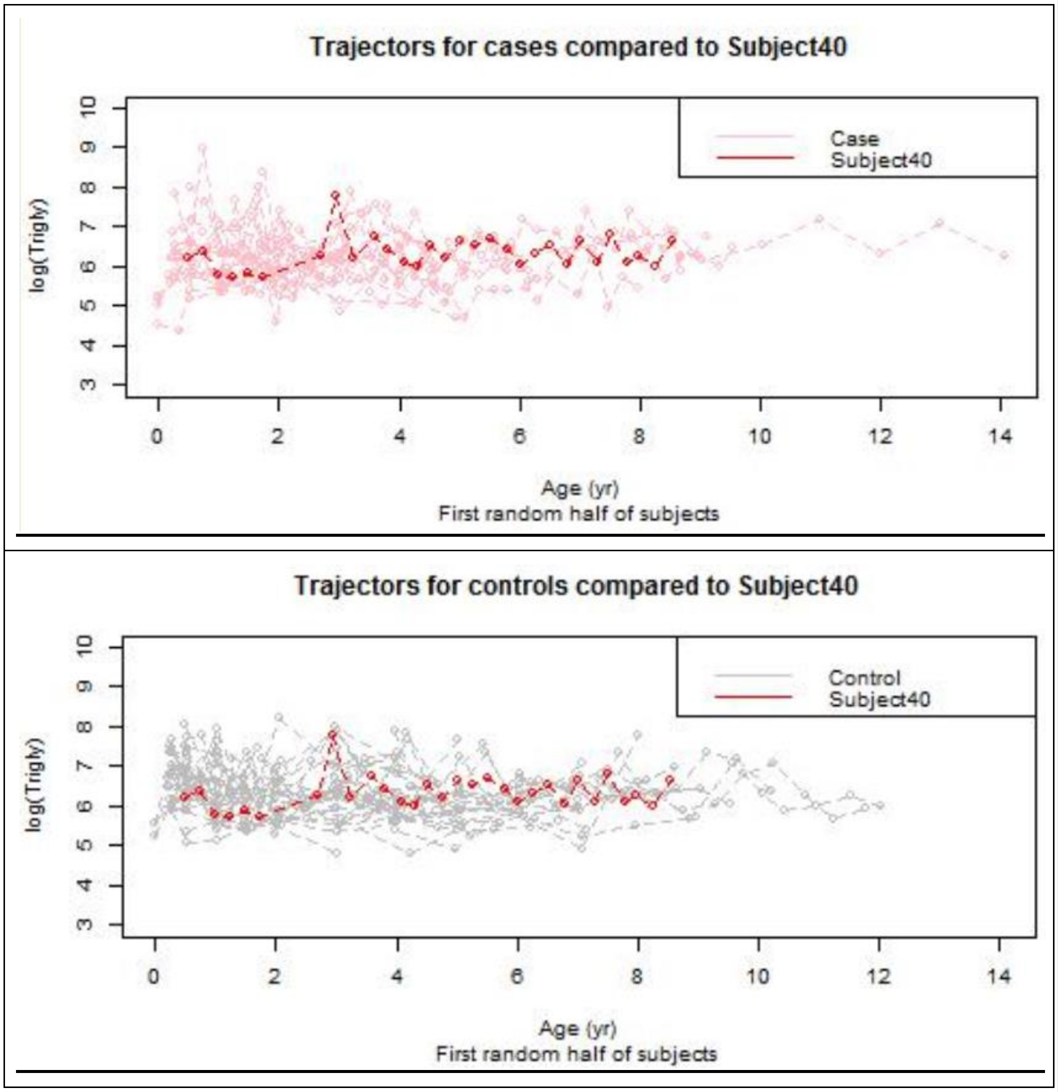


Table 3. EtherPC comparison (logarithm-transformed) using a linear mixed model

```
Fixed effects: log(Ether) ~ Case + AgeG

                  Value  Std.Error   DF  t-value p-value

(Intercept)  2.8228746 0.03887825 1016 72.60807  0.0000
Case1       -0.1134562 0.03080769   66 -3.68272  0.0005
AgeG2        0.0302731 0.02323005 1016  1.30319  0.1928
AgeG3        0.1318031 0.02616123 1016  5.03811  0.0000
AgeG4        0.1798098 0.02781331 1016  6.46488  0.0000
AgeG5        0.2115043 0.03041677 1016  6.95354  0.0000
AgeG6        0.2704272 0.03371040 1016  8.02207  0.0000
AgeG7        0.3275366 0.03714194 1016  8.81851  0.0000
AgeG8        0.2968349 0.04043353 1016  7.34131  0.0000
AgeG9        0.2980576 0.04771005 1016  6.24727  0.0000
AgeG10       0.2573967 0.05471428 1016  4.70438  0.0000
```

Lipid concentrations before seroconversion

This evaluation did not attempt to replicate the analyses of lipid concentrations before the appearance of autoantibodies. The abstract includes the following result: ”*Individuals who developed diabetes had… increased levels of proinflammatory lysoPCs several months before seroconversion to autoantibody positivity*”. Figure 5B in particular has been presented to support the claim.

It is true that there is a difference in one lysoPC [”*most abundant*”, PC(18:0/0:9)] between the cases and controls in the Mann-Whitney comparison. On the other hand, for parameter PC(18:2/0:0), the difference is in the other direction. For PC(18:1/0:0) there is no difference. The results for PC(18:3/0:0) are not presented, but based on Figure 2A of the article, it can be speculated that there would not be differences, as they did not appear in the age-group-specific comparisons. The analyses do not support the claim presented in the abstract of high values of lysoPCs (note the plural) before seroconversion.

Based on FDR, a majority of the differences with an observed significance level of below 5 per cent are real differences. It remains unclear on what grounds the parameters were selected for the comparisons: were all lipids in the data included in the comparisons, or were only some of the lipids selected for the comparisons?

4.3.2 Differences in metabolite concentrations

Replication

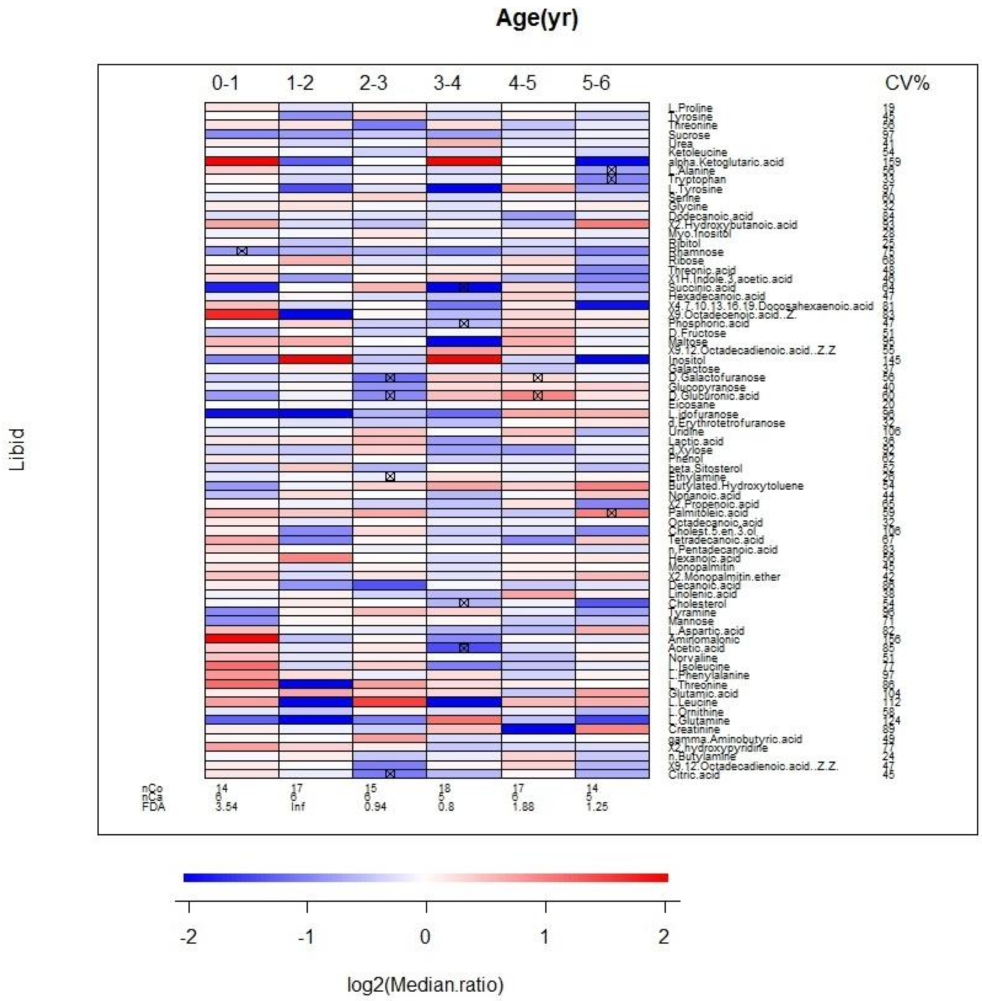
Metabolite concentrations were compared between the groups by repeating the analyses presented in Figure 4 of the article (with the exclusion of "birth", because the data was unavailable) using the same methods as in Section 4.3.1. The results are presented in Figure 5.

Taking all observations into consideration in the comparison, regardless of whether the observation was made after the age of diagnosis or at the corresponding age from the controls, gives results that are essentially identical with Figure 4 in the article.

The data set is small, and there are no visually clear differences evident between the cases and controls. There are few statistically significant results, and based on FDR they too can be considered primarily to have been caused by chance.

Figure 4 of the article also includes the comparisons made using placental blood samples. Out of the results, the result selected for the abstract was one that indicated that the succinic acid level of the cases is lower at birth than that of the controls. The reliability of the result is difficult to evaluate. Page 2979 of the article states that ”*succinic acid, a key metabolite of citric acid cycle, was diminished fourfold in progressors at birth and during the first year of life (P = 0.04)*”. In Figure 4 of the article, the significance level <0.05 is only for the placental blood sample; the low significance level has been extended to the first year's comparison, although the result was not significant. The generalisability of the comparisons at birth is questionable, as P=0.04 given on page 2979 presumably concerns precisely this comparison, and FDR for all of the comparisons is 0.16. These analyses were impossible to replicate, as the data at birth was unavailable.

Figure 5: Metabolite comparisons between cases and controls by age group, taking all observations into consideration



Metabolite concentrations before seroconversion

The findings according to which ketoleucine was low and glutamic acid elevated before the appearance of IAA and GADA were selected for the abstract of the article. This kind of trend appeared in the data, but the result should not be generalised: FDR is in the 0.7 range, which means that the rare differences for which P<0.05 can be largely assumed to be caused by chance.

4.3.3 Metabolism profiles

According to the preliminary analysis plan, Figures 2 A to C of the article would be repeated for all subjects. However, we decided to evaluate the most essential parameters one at a time, because it would have been difficult to make a reliable, quantitative evaluation of the simultaneous levels and variation of many metabolites while proportioning them to information on the times of seroconversion and diagnosis. On the other hand, as argued in Section 3.2, the differences that are reasonably permanent relative to time are the most useful for making predictions.

Based on the figures presented in this section, it can be visually evaluated

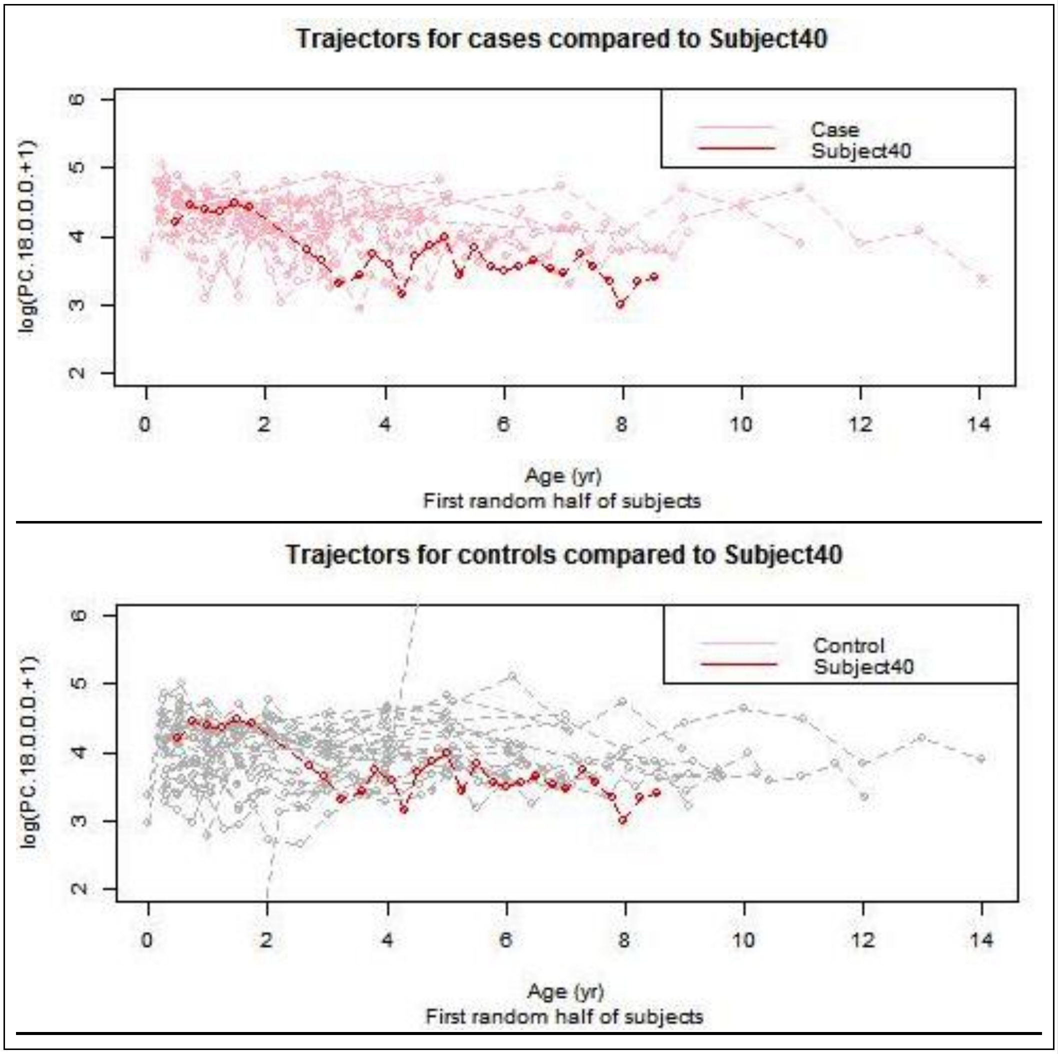
- whether the levels of the selected metabolites are higher for the cases or the controls;
- whether the high peaks in lysoPC concentrations are typical for people who have contracted T1D before the appearance of islet cell autoantibodies (ICA);
  - whether such peaks also occur for the controls;
- whether very low ketoleucine values and, on the other hand, high BCAA and glutamic acid values are typical for people who have contracted T1D just before the autoantibodies (IAA and GADA) appear.

In the figures the cases and controls are presented in separate figures. With regard to lipids, both groups are randomly divided into two figures in order to increase clarity. Case 40 is included in all figures. The results are presented using the natural logarithmic scale [ln (x+1)].

LysoPC

According to the article, the appearance of IAA was preceded by high concentrations of lysoPC. As was known, the values of case 40 were high during the first years and low after that in relation to others, but this applies both to the other cases and the controls. Overall, there appear to be no differences in the lysoPC levels between the cases and controls or the variations between persons or within persons.

Figure 6. LysoPC profiles for cases and controls

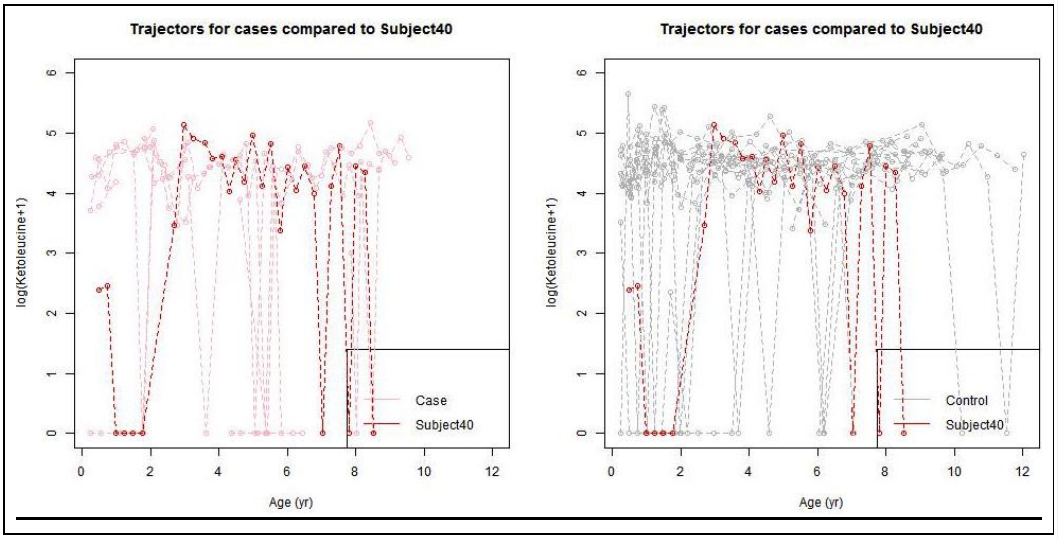


Ketoleucine

According to the article, positive IAA and GADA tests were preceded by low ketoleucine values.

The ketolecuine values in the data included several zeroes. Positive values were typically between 50 and 150. Visually, there is no difference between the profiles of the cases and the controls. The evaluator does not know how the zero values should be interpreted. The profile of case 40 does not appear different to the profile of a typical control.

Figure 7. Ketoleucine profiles for cases and controls

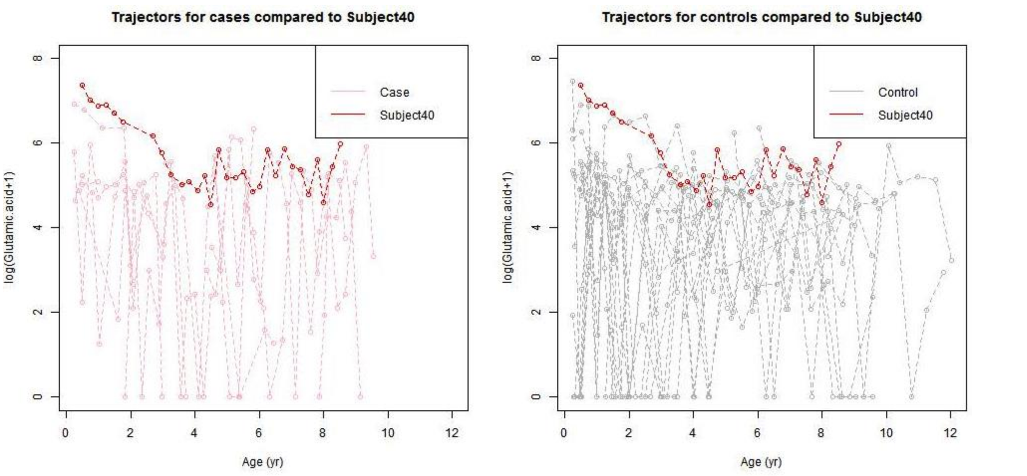


Glutamic acid

According to the article, positive IAA and GADA tests were preceded by elevated glutamic acid values.

The data includes several zeroes for this parameter, too, and the evaluator is unclear on how to interpret them. Large variations between time points are typical for both the cases and the controls, although the data includes children whose values remain relatively stable in relation to time. One of them is case 40, whose values are high compared to both the other cases and the controls.

Figure 8. Glutamic acid profiles for cases and controls

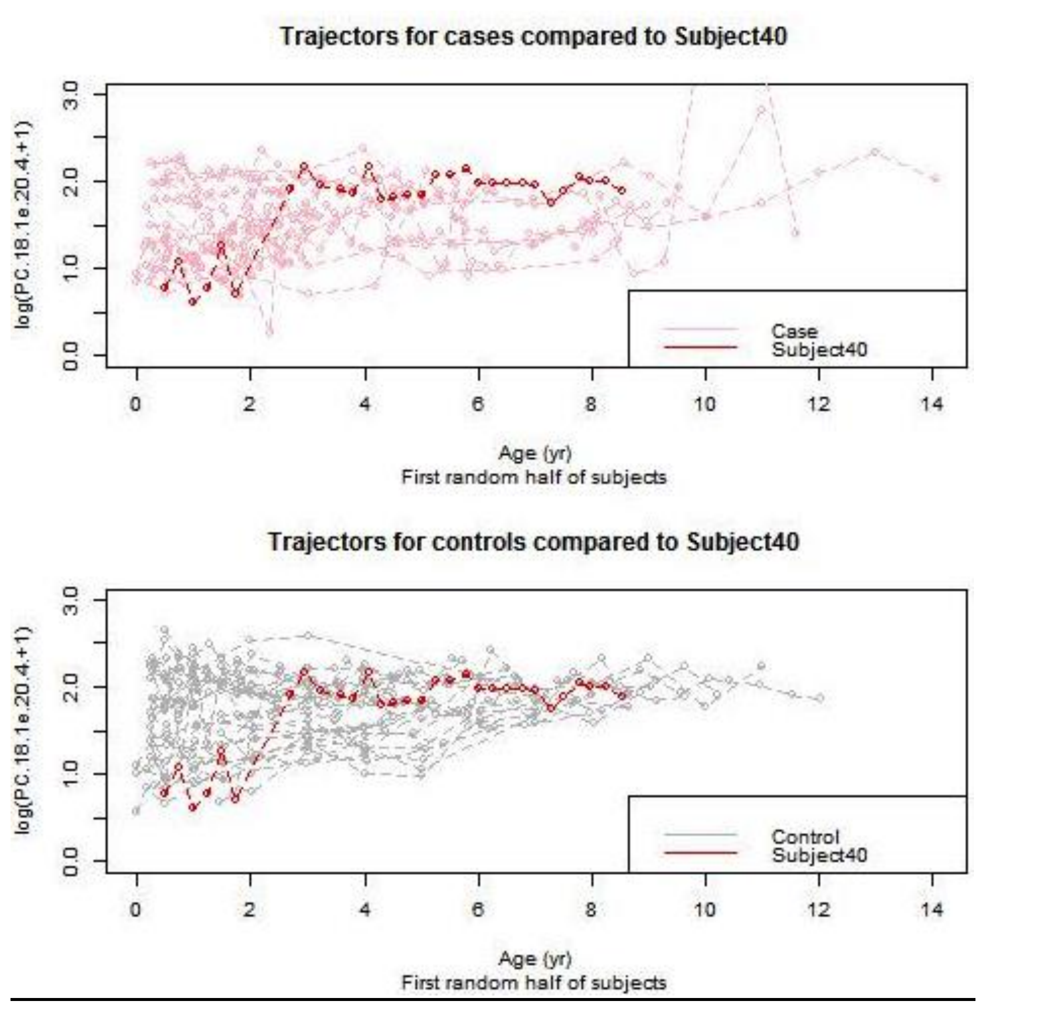


EtherPC

According to the article, the etherPC values of case 40 were abnormally low before the appearance of the first autoantibodies.

The figures in the Appendix confirm the claim that the etherPC values of case 40 were low until an age of about two years, after which they increased to elevated levels. The data of both the cases and the controls include corresponding increases between time points. The typical levels and variations do not seem to differ in the data of the cases and the controls.

Figure 9. EtherPC profiles for cases and controls

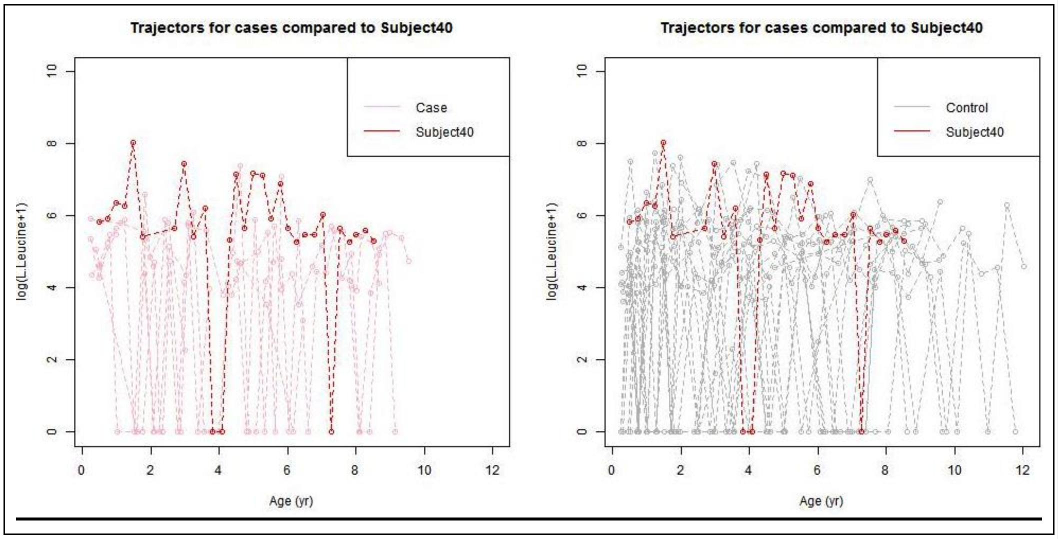


Leucine

According to the article, the leucine values of case 40 were abnormally high before the appearance of the first autoantibodies.

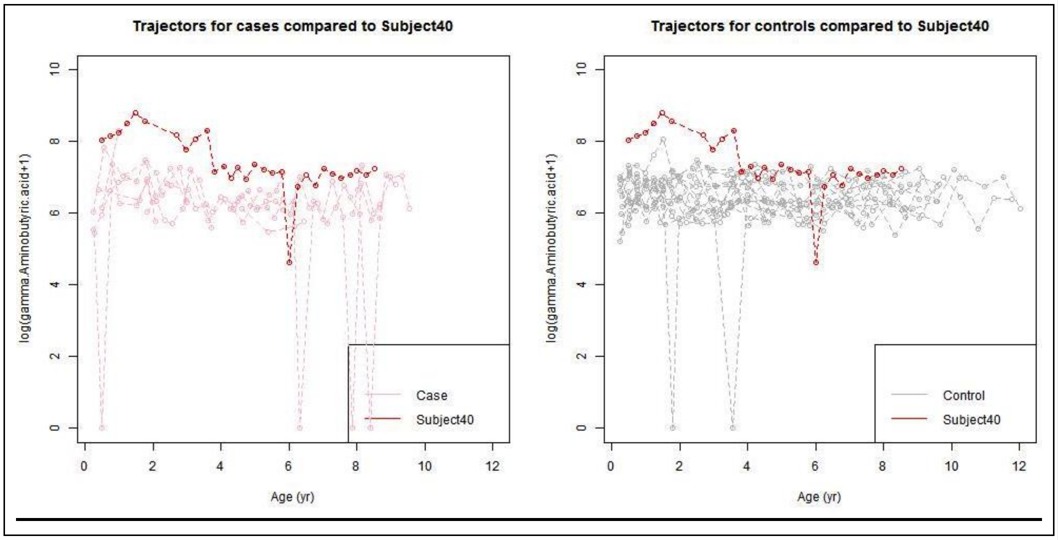
As seen in the figure (2B) in the article, the values of case 40 are mainly high compared to other children. This observation applies in relation to the other cases and the controls. The case has three measurements with zero as the value (interpretation remains unclear to the evaluator). There are also several peaks among the controls where the values are close to the 14-fold increase compared to "normal", separately mentioned in the article. Generally speaking, the profiles of the cases and the controls do not seem different.

Figure 10. Leucine profiles for cases and controls



Gamma-aminobutyric acid (GABA)

Figure 11. GABA profiles for cases and controls



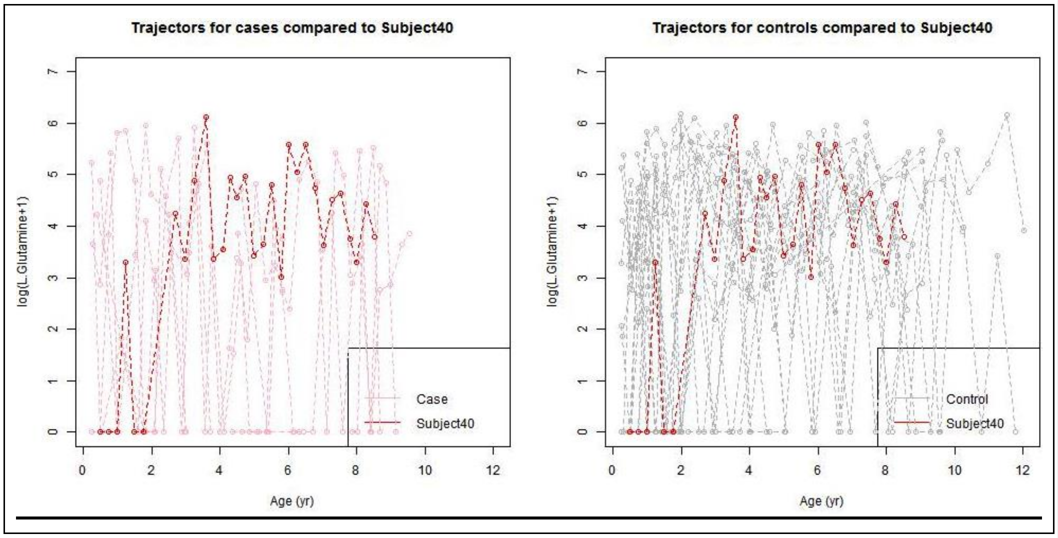
GABA values of case 40 are high compared to other children with the exception of one measurement. This observation applies in relation to the other cases and the controls. The general level and variation is similar in the cases and the controls.

Glutamine

According to the article, the glutamine values of case 40 were low before the appearance of the first autoantibodies.

The first values of case 40 were zeroes, and become positive only once before the age of two. After that, the values remained level, with variation that appears to be typical to both the cases and the controls.

Figure 12. Glutamine profiles for cases and controls

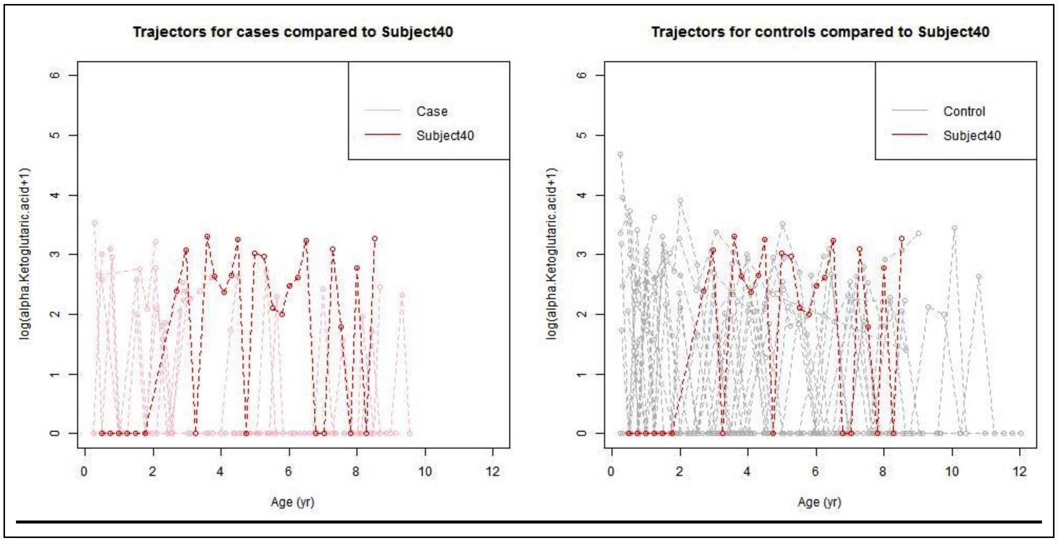


Alpha-ketoglutarate

According to the article, the alpha-ketoglutarate values of case 40 were low before the appearance of the first autoantibodies.

The values of case 40 were actually zeroes until an age of approximately two years, after which the values varied between "typical positive value" (around 10) and zero. The profile looks very typical compared to the cases and the controls.

Figure 13. Alpha-ketoglutarate profiles for cases and controls



<u>Summary of the metabolism profiles</u>

As a summary of the graphical comparison of the metabolism profiles, it can be stated that the claims concerning case 40 in the article are correct. The cases and the controls did not visually appear to be significantly different from each other with regard to any parameter. However, the values were examined relative to age in the figures without proportioning them to the seroconversion of each case. As stated before, this is not likely to be a large error, as only the age is a known variable for the predictive model, and for a majority the seroconversion took place before two years of age.

The large number of zeroes in the data was cause for consternation, as the article specifically stated that only those metabolites that were measurable in all samples were included in the analysis.

5 SUMMARY OF THE EVALUATION RESULTS

The results presented in the figures of the article can be replicated based on the available data with regard to the objective of this evaluation. The article has some slight imprecision with regard to the observations that have been used in the comparisons. Changing these details would probably not change the results significantly. Time of birth data was unavailable.

The statistical differences in triglycerides between the cases and the controls presented in the article appear reliable: the triglyceride and etherPC levels of the cases were lower than those of the controls. The differences were clearly statistically significant, but rather small compared to the other variance in the data.

The metabolomics comparisons presented in the article cannot be considered to be statistically reliable. Generally speaking, there are no statistically significant differences, and in my opinion, the article is in any case not particularly misleading in this regard. The results are presented as they are, although it is questionable to present results "P<0.05" in the figures with no corrections to the observed significance levels as a result of the FDR analysis. Because the data set is very small, results have been accepted as "statistically significant" even where their significance only emerges when the relative share of false findings is allowed to approach 100 per cent.

The article's text (e.g. discussion) does not clarify whether it is discussing the differences found in the data in a descriptive manner or presenting generalisable, statistically significant findings concerning the target population. The reader must speculate on the generalisability of the results based on the tense used or by perusing the detailed results. One cannot help but think that the differences between statistically significant and non-significant results have been purposefully obfuscated.

The examined metabolism profiles of sample case 40 appear to be abnormal in the case of some parameters, while others are typical among the other cases. Most essentially, the profiles of the cases and the controls do not visually appear to be different from each other.

The research results as such are not yet suitable for predicting seroconversion and T1D based on lipid values, because the observed differences were small in relation to the variance. My evaluation is that based on the article and the data, lipid parameters could be used as part of a predictive model for the onset of T1D, if the uncertainty related to the measurements could be reduced (e.g. by taking the measurements after fasting). The results thus provide information on the connection of lipid values and T1D and form a basis for further research and the possible development of a predictive model.

With regard to statistical methodology, the article may not be completely "by the book", but in my opinion, neither are there any clear errors. Indeed, in statistical analysis it is always relative whether an analysis has been performed "correctly", as all models are "incorrect" but some are useful. In my opinion, the article describes the statistical methods used to a sufficient – or at least regular – extent.

The evaluations did not even try to comment on whether the mechanisms presented in Figure 6 of the article are correct or even whether the data is in keeping with the figure in question. This would required knowledge of the field and other studies. Analysis of the data could be attempted using models, with the specific goal of determining the cause-effect relationships presented in Figure 6, but formulating such models would require preparations in cooperation with an expert in the field.

6 QUESTIONS THAT AROSE CONCERNING THE RESEARCH PLAN

1. As the article clearly states, the children had not fasted before the blood samples were taken. Can this cause large variations in the lipid values?
2. How is it possible that only 53 lipids ended up in the analysis? As a layman I would think that the same analysis method can be used to measure the same lipids, although some values could remain below a possible detection limit. What effect does it have on the results that only the predictiveness of positive lipids was assessed? (Quite possibly no effect, as not all samples have measurability that correlates with the differences between the groups.)