Falling at the Final Hurdle- Excluding Biostatistics at the Core of Experimental Design in the Field of Lipodomics

Mr H. Tong
School of Mathematics, University of Leeds, Leeds, LS2 9JT, UK

Dr J. Smith, Dr M. Zulyniak, Dr L. Marshall
School of Food Science and Nutrition, University of Leeds, Leeds, LS2 9JT, UK

Dr J. Wu
Dental Translational Clinical Research Unit, University of Leeds, Leeds, LS2 9JT, UK
(Dated: September 19, 2017)

ABSTRACT

Within the field of Lipodomics, complex techniques are often used to compensate for poor experimental design. It is normally the case that Scientists try to do too much in one experiment, resulting in noisy data that is difficult to interpret. This article will address the current state of the art in the field; comparing it to the foundations of the scientific method. It will discuss a major flaw in Lipodomics- the un-collaborative and fractured approach to research that sets the field up to fail, falling at the final hurdle- biostatistical analysis.

Keywords: Lipodomics, Biostatistics, Agile Methodology , Experimental Design, Machine Learning

I. INTRODUCTION

Conducting recent research in the field of Lipodomics has highlighted that the current way in which scientific research groups conduct their business is flawed. It relies on extremely advanced techniques, barely understood by the most competent of Mathematicians, to make sense of forever complex and large data sets. There is not one person in the research community to blame; however as technology has advanced, Scientists have become obsessed with these tools and overestimating their ability. Although the methods and techniques have become extremely sophisticated, structure needs to be maintained in the way in which projects are conducted.

This article will consider current and prospective techniques used by the Researchers. It will also consider the barriers met when working in a dynamic environment, such as biological research, and how certain methodologies can help overcome this. However, the major issue usually found in projects is that the basic principles of experimental design and the scientific method are often neglected. Consequently, many projects normally fall at the final hurdle- analysis.

II. CURRENT "STATE OF THE ART" IN LIPODOMIC BIOSTATISTICS

Before delving in to the techniques used by scientists in the field of Lipodomics, it is important to understand the nature of the information that experiments in this field produce- multivariate data. As the name suggests, this data consists of several variables that interact with each other in order to produce an effect that can be evidenced empirically. For example, if an individual consumes a multitude of different nutrients through their diet. The abundance of these nutrients can then be considered along with their effect on a persons health, measured through health indicators such as BMI and Cholesterol levels. Considering all of these variables will provide a much richer picture of an individuals health status than a univariate analysis.

Once this data is obtained, a lot of researchers in Lipodomics turn to a method called "Orthogonal Projections to Latent Structures" (O-PLS), a method whereby large data sets with uncorrelated variables can be processed such that the resulting analysis can be interpreted more easily. The algorithm attempts to model the underlying noise in the data to provide clearer data, regression is then applied in order to try and identify relationships in this data. This data can then be projected on to a 2D visualisation for cluster analysis.

The O-PLS concept is a primary tool used to consider the relationships in multivariate data and reveals interesting patterns of scientific significance. However, the data must be manipulated in a particular way for clustering to be meaningful, compensating for the poor initial design, that in turn produced sub-optimal data. O-PLS is an example of this manipulation, large and questionable adjustments are required in some cases to obtain the desired structured data to be analysed. Figure 1 explains how O-PLS works by use of a visualisation. Each variable is manipulated in to a vector and then these data points are projected on to a 'screen' in

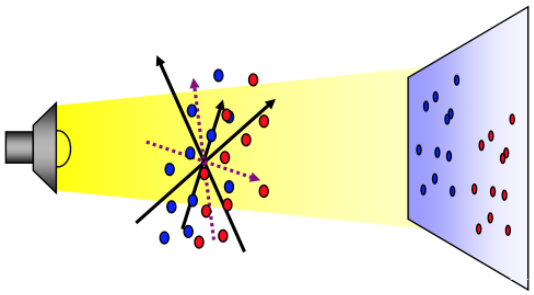

FIG. 1: OPLS explained [1]

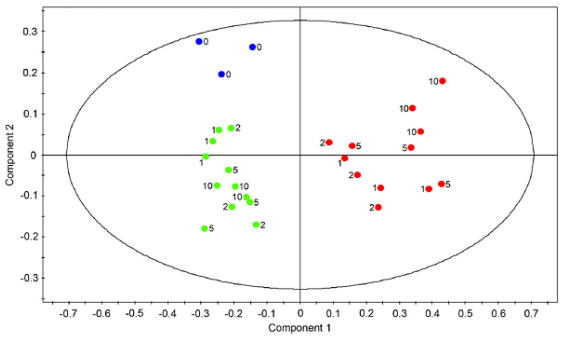

FIG. 2: OPLS example [2]

order to examine the relationship between these vectors. The impression created by the projection on to the 2-D surface so that it can be clustered and analysed is dubious at best. The "optimal angle" at which to project is calculated via a least-squares method for each example and is not standardised. Forcing the removal of noise may also render the analysis extremely difficult to interpret.

O-PLS is a multivariate regression analysis, an adaptation of the simpler PLS (Projections to Latent Structures). O-PLS is simpler to interpret than PLS, given an absence of noise and for a data set with a large number of participants- which is difficult to achieve. The technique is dependent on the fact that the compound and background noise are uncorrelated, otherwise it is no better than PLS. This being said, both of these techniques provide results that are hard to interpret in systems biology. This can be remedied by increasing the number of samples, this can be a time consuming and expensive process due to the nature of biological experiments. Caution is required to apply O-PLS and PLS to systems biology, which can be navigated by using preprocessed clusters and incorporating specific knowledge into whichever algorithm is being used.[3]

While the way in which the O-PLS technique works is not an issue for an experiment focussed on a qualitative question, where only the cluster structure is important; there is an issue with those centred around quantitative questions, especially those of a time dependent nature. This technique an angle that creates a 'best' result for the researcher that is the most likely to reveal a pattern, not one that is rigorous and universal. If the angle of projection were changed, this could reveal significantly change the interpretation of the experiment. [4]

Unsupervised Machine learning techniques have started to arise in the field of lipodomics, such as Hierarchical Clustering and Random Forest classification. These use different methods of grouping data, classifiers and clustering; however it is likely that one technique can provide different results to another. The random forest method is particularly prone to overfitting. In order for this method to work well, hundreds decision trees are required, with many data points. However, just as this may increase the precision of the classifier, it also presents the problem of false positives. The probability of significance then needs to be corrected in order to combat this error, the most common of which is the Bonferroni correction, which controls the False Discovery Rate (FDR). [5]

Regardless of how sophisticated these methods are, it is clear that there is a problem in the way in which they are executed if they yield different results. Hence, it brings in to question the original data collected from an experiment. Are scientists too focussed on using the currently available technology to its greatest capacity as opposed to collected the right data? Are researchers trying to answer too many questions at once with one experiment, under pressure to provide results with minimal resources?

At this moment in time, a lot of the research in Lipodomics is centred around clinical questions, as opposed to scientific studies. While these two types of investigation are related in that they use examine similar biochemical abundances; the way in which they are designed means that the data collected by Analytic Chemists about these structures is very different indeed. We will examine how the kind of question that a researcher is trying to answer governs their experimental design. Clinical studies usually examine cause and effect, observing the experiment from a more crude point of view. For example, a clinical lipodomic study might ask the question, "Does an increase in certain FA or Lipid create a particular effect in subject?". Using a combination of detailed meta-data and crude abundances of biochemical compounds from blood samples, before and after a meal, an intervention study can say with some certainty what causes a given effect. However, with a scientific and perhaps more theoretical study, one would need much more detailed data at a much higher granularity, in combination with a detailed meta-data profile of the individuals in question. It seeks to identify more subtle, and not necessarily observable, relationships such as the correlation

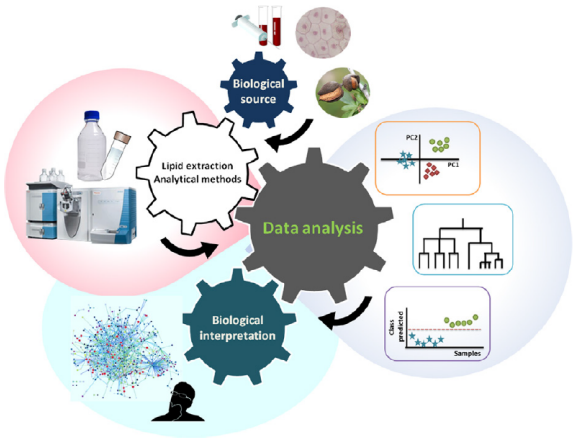

FIG. 3: Current thought in Lipidomic Data Analysis [6]

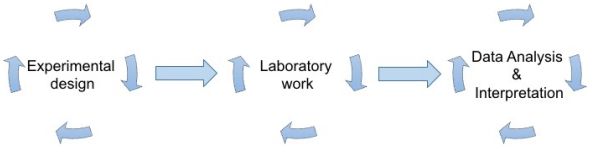

FIG. 4: Agile Bayesian Concept (ABC) Framework

between variables. For example, if one wanted to model a bio-chemical system graphically and probabilistically, it would be useful to have time-course data with lipid and fatty acid profiles, along information about an individuals health for stratification. Typically it is necessary for each strata to be greater than 30 in size in order to appropriately represent its population.

III. THE UNDERLYING ISSUE IS POOR EXPERIMENTAL DESIGN AND UNSUITABLE ANALYSES

A. Overcoming operational problems

In order to truly make any advances in the field of lipodomics, researchers need to work in more efficient manner whereby the aims of an experiment are periodically revisited in order to ensure the aims the group originally sets out to achieve are met. This includes integrating the data design with the experiment and iterating this process; ensuring that the data can be created with the analytical tools available and that once created, this data is suitable to be analysed for answering a specific question.

At this moment in time, biostatistics are mostly an afterthought in experimental design, they are not properly considered at the beginning and throughout the experiment. Perhaps the best practice for this is found in Software Development. The Agile movement works in small 'sprints' whereby an aspect of a project is iterated until it is completed to the highest standard. [7] The process includes regular daily short meetings to update the team and for collaboration, to ensure the 'sprint' is efficient as possible. For example, a project could be designed as follows.

1. At the beginning of an experiment, a research group would iterate over the requirements of a project, taking in to account the equipment and tools required. Individuals can properly research the area and techniques before deciding on a method. In research this may require the researchers to adapt, modify or invent new methods in order to properly meet the needs of their problem. After this, the group would create a proof of concept (P.O.C.) whereby the experimentalists conduct a trial run with the chosen techniques; the academics modelling the processes will also be involved at this stage to ensure that the work produced meets their requirements. Hence, after this the project will loop back round to the research and design stage or progress to the next 'sprint' of conducting the experiment, ensuring all parties are integrated in to every stage of the experiment.

2. The experimental phase would be lead by those conducting the experiments, to ensure that this step is executed efficiently. Statisticians, Data Scientists and more theoretical scientists would then audit this phase to ensure the experiment meets the objectives and help navigate any unforeseen complications. Not only will this significantly enhance the efficiency of the project; those analysing the data will have a far greater context and understanding, which will in turn result in a much more holistic review of the subject at hand. This sprint can be split in to several sprints if the experiment is large with many sections

3. The analytical stage will now be much more efficient than in an regular academic project. Since statistics resonates throughout the whole of the project, essentially governing how it is carried out, then so long as the data complies with the requirements then little manipulation will be required in order to achieve the desired results. In effect this is where the role of the laboratory and theoretical scientists switch, those who understand the underpinning biochemistry and physical concepts will govern the analysis. The statistician should take a Bayesian approach, using empirical chemical concepts as a prior to inform their investigation. This will become particularly important for experiments conducted with a large number of data points. From a computational stand point, the data

will be a lot easier to digest if it can be broken down in to smaller groups, particularly when using unsupervised machine learning techniques. The concept of subsetting the data is extremely useful in biological experiments and will be explored in the next section.

B. Overcoming technical problems

Analysing the data produced by an experiment is the final and perhaps most important stage of any scientific project. It is the information that evidences the work of the group and allows researchers to come to conclusions and perhaps unexpected conclusions. We have already certified why it is important to make sure the data is considered throughout for operational purposes; it is then important to consider the different types and formats of data in order to make the project as rigorous as possible. One cannot reiterate how important it is to have an appropriate data appropriate to the problem you are trying to solve; if the data is inadequate it will render all previous work useless.

First, however, we should consider the fundamental mathematical concepts that require the data to fulfil certain requirements. The Central Limit Theorem requires the sample size, n, to be "sufficiently large" in order to approximate the distribution of the population, this is usually n$\geq$30. With biological studies, this is usually an issue, as the number of variables often outweighs the sample size. In a lipodomics concept this would mean having at least 30 readings of each chemical abundance so that a distribution can be estimated. It is then interesting to use a time series analysis on samples collected at regular time intervals, e.g. $\Delta t = 30$ minutes. In this case an intervention study would be appropriate, whereby the subjects fast for 24 hours before the test and then eat a meal, with blood samples being taken every 30 minutes for 2 and a half hours. However, this may bring up some ethically issues- which is why a diverse team of individuals is necessary to mitigate these risks.

An intervention study is extremely useful when using biological data as it minimises the number of variables needed to be controlled in the experiment, it allows a Mathematician to make solid assumptions when considering the data. In this type of study, the underlying biological activity can be assumed constant for the whole experiment and so the only thing affecting the abundance of the biochemical compounds concerned, can be assumed to be the food. Since the abundance of these compounds taken in through food and transferred into the blood are known, the reaction rates between different molecules can be inferred.

The sample size of the overall population is not just governed by this relatively light constraint, but by the meta data which is collected from the subjects. BMI, age and gender all act as variables in biological studies and can easily be standardised by stratification. However, it is also necessary for each of these strata to be of size $n \geq 30$ to form a distribution for each of their respective populations. Hence, the total sample size will need to be of size m where $n \geq 30m$. This will not only provide insights in to the difference between naturally occurring groups; but also identify any other complications, such as noise that might distort otherwise clear relationships in the data.

Another important consideration, mentioned earlier, is the Bayesian approach to an academic project. Within research teams there are experts on the subject matter at hand and so it would be a waste not to include their opinions to help form conclusions. Particularly in Biochemistry, there are rules and exceptions for reactions that are universally known. This presents a combinatorics problem, whereby a mathematician can match up all possible reactions in their data set, in order to minimise the amount of data needing to be processed. For example, one Fatty Acid can only combine with a finite number of other Fatty Acids in order to combine to form a Phospholipid. Although it is important to also consider all the relationships, to ensure nothing is missed, this method will ensure computation is significantly more efficient. As such, the grouping can be isolated and examined in much greater detail.

C. Primary Analytical Techniques

Now that we have everything in order, we have a data set that is ideal for the problem we are trying to solve the final step is to choose an appropriate analytical method. Conducting couple of different methods would also be ideal in order to verify the results. In order to do this, it is important to survey the whole data set with quick and computationally inexpensive techniques. The obvious choice here is to perform a descriptive investigation on all variables. Histograms of each variable are also useful in order to visualise how it is distributed. Due to the ideal nature of the data collected, stipulated and governed from the beginning of the project, by $n \geq 30$, by the Central Limit Theorem these variables will be normally distributed with mean μ and variance σ^2. The interesting discoveries come with modelling how these distributions change over time.

All the techniques mentioned so far only take each variable in isolation, however. In order to understand the biochemistry of lipodomics in more depth, it is necessary to examine the interaction between different biological compounds. For example, how one fatty acid reacts with

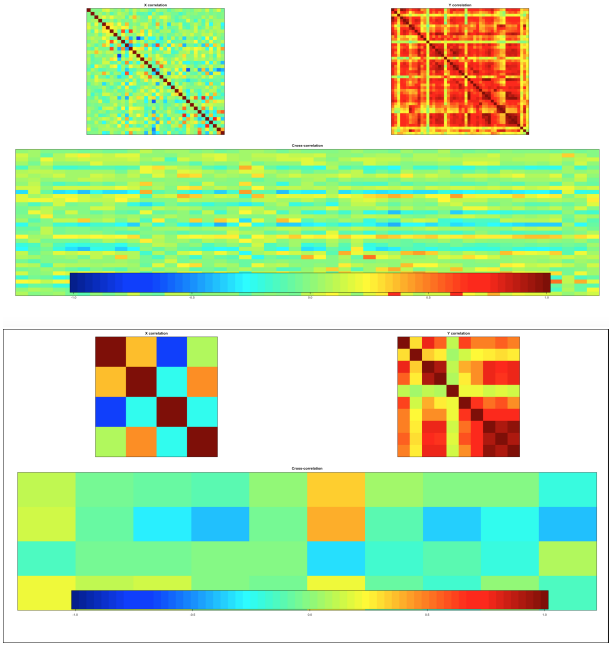

FIG. 5: CCA Correlation Matrix Plot with **a)** Whole data set **b)** One Strata of the data set

other fatty acids to for phospholipids and triglycerides. A useful primary technique for this is Canonical Correlation Analysis. This can be rigorously defined as follows.

Canonical correlation analysis is method for exploring the relationships, specifically the correlation, between two multivariate sets of variables, in this case the fatty acids $\left(X_{FA}\right)$ *and their corresponding lipids* $\left(Y_{Lip}\right)$*. This can be defined mathematically as follows.*
Let X_i *for* $i = 1, ..., k$ *be the fatty acids and* Y_j *for* $j = 1, ..., n$ *be the lipids, then:*

$$\underline{X}_{Nxk} = \underline{Y}_{Nxn}\underline{\rho}_{nxk} + \underline{E}_{NxK} \sim MVN(0_{N.k.}, \underline{\Sigma}_{kxk}) \tag{III.1}$$

Where '$\underline{\rho}$*' is the correlation between the two data sets and '*$\underline{E}$*' is the bias.* [8]

As shown in Figure 5, there is a noticeable benefit to prototyping the technique on one strata of the data. In the context of lipodomics, the top left square shows the intra-set relationship between the Fatty Acids and the top right square

D. Advanced Analytical Techniques

Once the primary analysis is complete, we can now start to use more sophisticated techniques to uncover more detailed information about the relationships between our variables. This includes using unsupervised machine learning techniques and modified Monte Carlo methods as opposed to Monte Carlo methods in order to optimise the research. In simple terms, unsupervised machine learning methods look for patterns in unlabelled data, using statistical methods, such as Maximal Likelihood. This is opposed to semi-supervised or supervised machine learning techniques, whereby labelled or partially labelled data is used to train an algorithm to predict further outcomes by generating a 'typical' or expected outcome for an individual input given a specific set of variables. An example of this is that an Obese individual might have a specific lipid or fatty acid profile; alternatively, an individual with Addison's disease might have another lipid or fatty acid profile. This is more useful in diagnostic or clinical investigations; however here we shall discuss 3 unsupervised methods, Bayesian Hierarchical Clustering (BHC), Gaussian Mixture Models (GMMs) or 'Fuzzy Clustering' and Hierarchical Dirichlet Processes (HDP): since the article focusses on a more scientific and investigative approach. These techniques should be trialled on one strata only for testing purposes to ensure that a technique is appropriate, as conducting the procedure on a whole data set may be time and computationally expensive. It should be noted that there are many different types of machine learning algorithms and computational packages. A research team should thoroughly investigate and consider all the options for their project before starting. This might include choosing several techniques to compare the results; and consulting professionals from other industries, such as Finance and Computing, to see if there are options that a team might not necessarily consider.

First, let us consider Bayesian Hierarchical Clustering. [9] This technique is an attempt to speed up the Infinite Mixture Models, outputting a hierarchical structure such as in figure 6. The algorithm first runs a script to 'find optimal binning', a process whereby the data is discretised and clustered in to three groups, 0, 1 and 2, which are dissections of a normal distributions. Once the optimal ratio for the size of each 'bin' or group is found, the euclidean distance between all clusters is found, which can be plotted as a dendrogram. The tree is organised in such a fashion whereby the lower 'leaves' represent the original data, then as one traverses up the structure, these clusters merge until there is only one cluster remaining. This reveals how closely related certain variables are, compared to the rest; whereas CCA only presents pairwise relationships. This method should be used instead of Markov Chain Monte Carlo (MCMC), since it is more computationally efficient taking significantly less time to run. However, for a substantially large data set, it might be worth using a super computer. For more details and precompiled code, one should consult the 'BHC' package which can be found in the BiocLite repository on the Bioconductor, using the below code in R statistical. For a more in depth and mathematical explanation of this method,

```
source(''https://bioconductor.org/biocLite.R'')
biocLite(''BHC'')
```

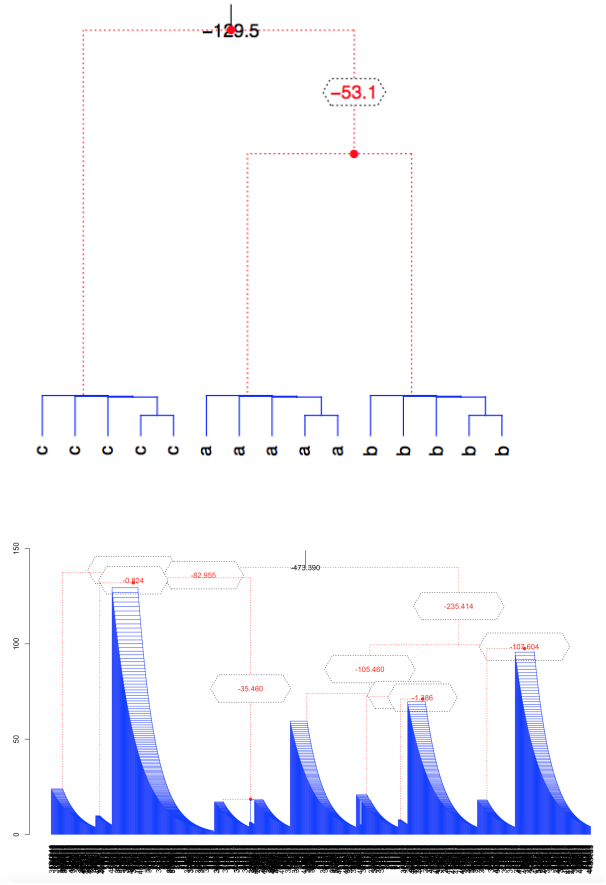

FIG. 6: BHC hierarchical tree output for **a)** A small data set **b)** A large, more complex data set

Next, we consider a non-parametric method called Gaussian Mixture Models (GMMs) or 'Fuzzy Clustering'. [10] In the machine learning community, Gaussian Processes have often been viewed as a 'black box' whereby a researcher is not concerned by the mechanics of the method, so long as it makes good predictions. Rasmussen succinctly summarises this area of analysis by describing it, as below.

"The hierarchical formulation of the covariance functions with hyper-parameters, the testing of different hypotheses and the adaptation of the hyper-parameters gives an excellent opportunity to understand more about the data" [10]

In lay terms, this method optimises the gaussian distribution of a variable, using a tool called a hyper-parameter as an indication as to whether this is the optimal representation of the data set. These distributions reveal interesting features of a variable, such as why the data set has a particular covariance structure. In a lipodomics setting, this would be how the abundance of a biochemical compound is distributed and how this might affect the relationship between different compounds.

Finally, we will discuss a Hierarchical Dirichlet Process (HDP) approach. [11] This builds on the GMM approach, but allows us to tie the mixture models in to different groups. The method is based under the assumption that the number of mixture components is unknown a priori; however, the mixture models that share a group necessarily share these mixture components. Hence, since these models are separated in to groups, we can consider a hierarchical approach, whereby the measure for child Dirichlet process is governed according to a different Dirichlet Process. This representation leads to an MCMC sampling scheme for posterior inference. Following a hierarchical, Bayesian, non-parametric approach has already proven useful in Bio-informatics to advise on the uncertainty surrounding the approach number of clusters in these data structures.

All the above techniques take a Bayesian approach, which is well appreciated by the machine learning community, but not necessarily by the statistics community (Rasmussen, 2009). The reason for this, as mentioned before, is to minimise the number of variables and hence optimise the analysis for computational efficiency. Especially in the field of Bio-statistics, data sets can be extremely large and if procedures like this are not followed, the analysis can take an inconceivable amount of time to process.

IV. A CASE STUDY- AN ANALYSIS OF THE RELATIONSHIP BETWEEN FATTY ACIDS, PHOSPHOLIPIDS AND TRIGLYCERIDES

Before writing this article, the research group set out to answer a very specific question in Lipodomics, through an EPSRC 10 week scholarship at the University of Leeds. The question considered how a pool of 10 Fatty Acids combine together to form lipids. This was originally designed to look at the combinations that occur to form phospholipids (lipids with two fatty acid tails); before moving on to consider the more complex problem of triglycerides (lipids containing 3 fatty acids). This investigation is unique in that Lipodomics only usually considers lipid profiles as a means to relate them to a condition; not in their own right, which would provide a more canonical and useful understanding of this important part of the metabolism.

In order to investigate the relationship between these biological compounds, we had to find suitable data to answer the questions. However to our disappointment the major studies for the relevant compounds were not suitable for the research that the group wanted to conduct. We considered three pieces of data, detailed below, before

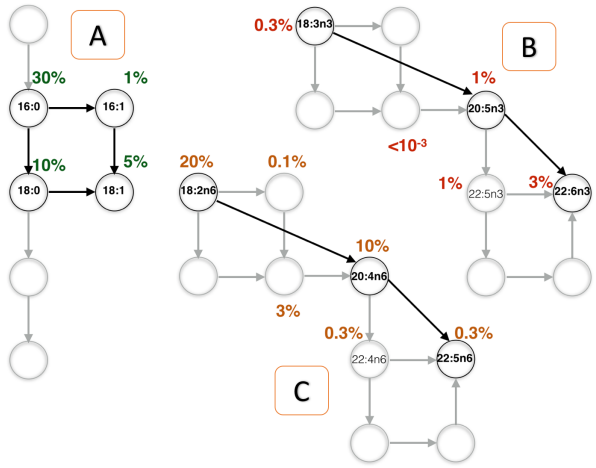

FIG. 7: A Directed Acyclic Graph (DAG) showing the relationship between 10 Fatty Acids. The percentages indicate the relative abundance (or ratio) of each fatty acid.

conducting some primary and more advanced analysis on the data sets. This includes:

1. European Prospective Investigation into Cancer and Nutrition (EPIC) study including 46,000 participants and their fatty acid profiles.

2. National Diet and Nutrition Survey data with profiles of fatty acids, phospholipids and triglycerides.

3. An intervention study whereby individuals were fed a Beef burger meal from a fasting state.

The first two data sets were not suitable for many reasons, but the most pressing issue was the fact that the readings were only taken at one time point. This is important to establish the rates at which the fatty acids react together to form lipids. The second data set was better than the first, although not ideal, since it had abundances of all compounds. The data sets could not be used in tandem since they were not studies on the same individuals. This could not be generalised in to strata either, since there was missing meta data for one of the sets. Ultimately this did not matter, as the time dependent nature of the data is vital to revealing anything interesting about the compounds.

The intervention study was then found, this included meta data, had abundances for all chemical compounds and was time dependent. The only thing letting this study down was its size. Disregarding incomplete data, the study only had $n = 29$ admissible participants. Hence, although all the information about these individuals was present to categorise them in to strata, this could not be done. With groups of this small size, any information could not be generalised to a population with some arbitrary common characteristic, by the Central Limit Theorem.

Given that the research group had been given significantly more time and funding, we would have used the above framework to ensure that the ideal resulting data had several key features in order to meet our objective. This is listed below in order to act as a non-exhaustive checklist, to ensure all criteria is met by the project.

1. An ideal data set would be of a time dependent nature with a consistent Δt (perhaps at 30 minute intervals). This allows for rates of reaction to be calculated; and such the ability to compare biological relationships in more depth.

2. Stratification of data for noise reduction and examining how ones findings differ throughout the population. For example, if a group of individuals have low abundances of Fatty Acids present in breast milk could indicate that they require growth factors.

3. It is necessary for each strata to have at least 30 individuals in order to accurately represent all individuals in the population with a common characteristic.

4. Intervention studies are suited to studies in lipodomics to account for background biological activity. For example, if a group was fed a Beef burger meal, they would expect to have a higher abundance of the FAs in DAG C in figure 7; whereas those fed a meal such as fish, rich in omega 3 and 6 oils, they would expect to have a higher abundance of FAs shown in DAG B.

5. Neatly labelled variables and detailed notes on the way in which the experiment was conducted. Not only will this make the analysis more efficient; it will also result in the final report being more coherent.

6. A list of physically possible chemical combinations of fatty acids to be used in the analysis and report. This may also make experimental work more efficient, as well as the analysis.

Once the team acquires the appropriate data, we approach the final hurdle- analysis. From an early stage of this project, the group knew the exact framework that we would have liked to have used to make inferences from our investigation. This started with realising the type of data we have acquired and asking ourselves which way in which would be most suitable. Since the question was addressing an area of Lipodomics that has not been investigated before, the data was 'unlabelled' with respect to the biological interactions. Hence, this was most suited to unsupervised machine learning techniques on the correlations between the fatty acids, phospholipids and triglycerides. This is whereby we know nothing about these relationships or the correlation between the compounds, so it is necessary

to discover this using techniques such as clustering. In this case the previously explained 'BHC' would be the most suitable. Not only does this cluster the relevant bio-chemical compounds in to groups that are most likely to combine to create certain lipids, it provides structure. This structure gives a quantitative insight in to how closely all considered compounds are related, with regards to combining to create lipids.

Whilst the group will not specifically use the meta and qualitative data for analysis intrinsically, it is still important. The meta data allows the individuals to be split in to strata, which is useful for a couple of reasons. Firstly, it will allow comparison to see if the groups findings are canonical, or specific to one group of people. For example, a clinically underweight strata may require a lipid that promotes growth much more than an obese strata. It will also remove any 'noise' present in the data, allowing the analysis to be much easier to interpret. Since individuals in a strata are much similar to each other than the population, there will be a greater number of standardised variables- which means less variables to skew ones findings. The number of variables can also be reduced by using qualitative information from chemists, as mentioned before. Some chemical compounds simply cannot physically combine to create certain lipids; hence we can exclude these from the analysis. If required the compounds can be further split in to smaller known groups for more concentrated analysis. As can be seen in figure7 , the fatty acids are naturally split in to three directed acyclic graphs. Fatty acids in the left hand DAG can be analysed in combination with either of the other graphs, or in isolation. Hence, it is wise to analyse the DAG containing the fatty acids C16:0, C16:1, C18:0 and C18:1 first to test the framework on a group that are empirically known to combine to create common and useful lipids.

In terms of the analytical techniques used, the framework is systematic. The methodology that the group would use has been laid out in a list format below. This is to reiterate the fact that if one step of the process is unsuccessful, then the rest of the techniques are rendered inadmissible. This should be treated as an algorithm for consistency; however one should realise that this 'final hurdle' should not be as a hurdle at all if the data meets their requirements from the very start. An Agile and iterative methodology will help to ensure this, which will lead to the group not wasting time and resources.

Analysis framework for case study

Given that the previously laid out agile framework had been followed and appropriate data has been collected, one would work through the following steps.

1. Descriptive statistics for each of the fatty acid and lipid abundances. *(Their abundances can also be modelled accurately using an infinite dirichlet mixture model, not outlined in this article)*
2. Canonical Correlation Analysis to investigate the relationships between all variables. *(This step can be revisited further on to ensure that more complex results agree with this primary analysis)*. This step is more qualitative than quantitative– it provides information on how strong the relationships are but not anything more than that, such as rates or probabilities.
3. Bayesian Analysis – Use combinatorics (empirical information from physical chemistry) as a prior in order to minimise variables by which fatty acids can react with each other.
4. Bayesian Hierarchical Clustering– A more advanced analysis to understand the relationships between different compounds and how these are structured, with a focus on how strongly each fatty acid has an affinity to react with another.
5. Hierarchichal Dirichlet Process approach– A different machine learning approach that should consolidate with both the BHC and CCA. This instead focusses on the distribution of each variable. This can be useful to infer probabilities in the DAGs pictured in figure reffig: FAs
6. Inference– Using both the qualitative analysis and qualitative information conclude insights in to the way in which the variables interact. Consider the whole population as well as each strata, as the data could have clinical implications alongside the scientific discovery.

V. A NOTE – THINGS TO CONSIDER AND BE AWARE OF WHEN USING UNFAMILIAR TECHNIQUES IN BIO-STATISTICS

This final section mentions a few things that individual might find useful if they are new to the field; but equally to reiterate for each project. Some are unique to Lipodomics and affect the way in which ones findings should be interpreted; other points could be relevant to the wider scientific community.

Lipodomics in particular is interesting due to a paradox which highlights the fact that if a biological compound is in low abundances, this means that it is in high demand. This is counter intuitive and can lead to issues whereby there are naturally low abundances in a particular biological molecule and/or background reactions occurring within the body. As previously mentioned this can be combatted by an intervention study, whereby the participant is fasted for a period before being fed a specific meal. A time dependent data set will also provide higher granularity, to monitor the levels

of a compound very closely and spot any unexpected behaviour. Stratification has been previously mentioned, mainly for the fact that it factors in greater controls and allows the group to verify whether the conclusions made are canonical or not; however it is also useful to make the process more efficient. In this final 'sprint' of the agile project, it is quicker and easier to prototype on one strata before writing code and analysing much larger and more complex data sets.

Within science, it is valuable to hypothesise about the results that you should expect to see. Hence, the prior qualitative information used to reduce the number of variables by the physically possible reactions between compounds can be extended to verify ones conclusions, ensuring that the qualitative and quantitative information agree. An example of this in Lipodomics (previously mentioned) is related to the results expected from the different meals fed to an individual in an intervention study. From previous experiments, we know that if a study was fed a beef burger, one would expect FAs in DAGs A and C to dominate; alternatively if the individual had a meal of fish, one would expect the FAs in DAGs A and B to dominate. This checks that the abundances are as one would expect before moving on to conduct more interesting analysis about their reaction rates. Alongside the quantitative checks of ensuring all quantitative techniques align, the project will present rigorous evidence of its findings.

In this article we have talked a lot about finding the right techniques and methodologies for your project. Many of these you might have never heard of before, but a simple search of the internet will reveal simple implementations and best practices for your chosen technique or methodology. For example, if one were looking to implement the Agile methodology, there are plenty of courses and papers available, such as PRINCE2, that will ensure that this is used properly. [12] From a more technical standpoint, if one wanted to train an Artificial Neural Network (ANN) to diagnose a disease from a lipid profile, there are parallels in Finance with banks using supervised machine learning to identify fraud from an individuals bank statements. [13] This might seem extremely complicated; however using pre-compiled packages, along with their vignettes and previous use cases, one should be able to efficiently use any technique. It is important to understand that these resources must be used with caution. The code for these techniques in languages such as R Statistical and Python can be extracted from a package. This code should then be checked and modified for the problem that you are trying to solve. If there is any difficulty with this one should consult an expert in your institution or online.

Finally, it is important to highlight the compromises that may be necessary with real world problems. Although we have always said that the quality of the data is paramount, there may be obstacles to this from a practical and ethical standpoint. It might not be physically possible for a researcher to complete the experimental phase in the way that the problem requires, at this early point the team should iterate over different possibilities and come to an alternative solution. Equally, it may not be ethical for the experiments to be conducted in a certain way. A perfect example of this is when using time dependent data, at what time interval is it ethical to take blood samples at? The ideal abundance readings would be continuous, instead of at discrete time steps, which is not practical. Hence, there is a trade of between the smallest Δt for precise readings and what is permitted by ethical bodies.

VI. CONCLUSION

If there is one thing that you should take from this article, it is to keep it simple. An individual within a diverse team should know their limitations, the limitations of the data and the limitations of the techniques they are working with. The experimentalists should not focus on maximising what the pieces of equipment they have are capable of, but using these powerful resources to produce the most suitable data. They should seek advice from other members of the team and community about matters that are not within their skillset, coming to an understanding with the more theoretical team members of how the project should move forward. This might involve compromise between the Statisticians and Analytical Chemists, as in the above case study. Those focussed on the analysis should in turn collaborate with the experimentalists, realising that the data alone does not tell the full story. Using previous studies and expert opinions will allow the team to work quicker and more efficiently, not wasting any resources, such as funding and expensive equipment.

Keeping your project simple, yet scalable should not be just seen as an operational efficiency exercise, but also as a quality assurance practice. Simplicity allows the team to monitor and control the whole environment of the experiments and hence reducing noise and error in the process. As a result, the group will produce clearer results that have been acquired through less computation; these findings will be rigorous, evidenced by both qualitative and strong qualitative analysis.

Acknowledgements

We thank Albert Koulman for providing the data sets that were used to prototype the case study for this article. We also thank EPSRC for funding.

[1] Susanne Wiklund. Multivariate data analysis for omics, 2008.

[2] Henri S. Tapp and E. Kate Kemsley. Notes on the practical utility of opls. *TrAC Trends in Analytical Chemistry*, 28(11):1322 – 1327, 2009.

[3] David J. Biagioni, David P. Astling, Peter Graf, and Mark F. Davis. Orthogonal projection to latent structures solution properties for chemometrics and systems biology data. *Journal of Chemometrics*, 25(9):514–525, 2011.

[4] Johan Trygg and Svante Wold. Orthogonal projections to latent structures (o-pls). *Journal of Chemometrics*, 16(3):119–128, 2002.

[5] Richard A. Armstrong. When to use the bonferroni correction. *Ophthalmic and Physiological Optics*, 34(5):502–508, 2014.

[6] Antonio Checa, Carmen Bedia, and Joaquim Jaumot. Lipidomic data analysis: Tutorial, practical guidelines and applications. *Analytica Chimica Acta*, 885:1 – 16, 2015.

[7] cPrime. What is agile? what is scrum? https://www.cprime.com/resources/what-is-agile-what-is-scrum/ Accessed 14th September 2017, 2017.

[8] Ignacio Gonzalez and Sebastien Dejean. *Package 'CCA'*. CRAN R Repository, 1 edition, 2015.

[9] Katherine A. Heller and Zoubin Ghahramani. A nonparametric bayesian approach to modeling overlapping clusters. In Marina Meila and Xiaotong Shen, editors, *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, volume 2 of *Proceedings of Machine Learning Research*, pages 187–194, San Juan, Puerto Rico, 21–24 Mar 2007. PMLR.

[10] Carl Rasmussen and Christopher Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2 edition, 2006.

[11] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[12] Prince2.com. prince2 information prince2 courses for project managers provided by ilx group. https://www.prince2.com/uk/what-is-prince2, 2017.

[13] Jacomo Corbo, Chris Wigley, and Carlo Giovine. *Applying analytics in financial institutions' fight against fraud*. QuantumBlack, 1 edition, 2017.