Over recent years many statisticians and researchers have highlighted that statistical inference would benefit from a better use and understanding of hypothesis testing, $p$-values, and statistical significance. We highlight three recommendations in the context of biochemical sciences. First recommendation: to improve the biological interpretation of biochemical data, do not use $p$-values (or similar test statistics) as thresholded values to select biomolecules. Second recommendation: to improve comparison among studies and to achieve robust knowledge, perform complete reporting of data. Third recommendation: statistical analyses should be reported completely with exact numbers (not as asterisks or inequalities). Owing to the high number of variables, a better use of statistics is of special importance in omic studies.

**To retire, or not to retire, statistical significance, that is the question**

Despite their ubiquitous application, the current use of statistical significance (see Glossary), $p$-values, and other test statistics is a source of concern and controversy among statisticians and researchers. In 2016 the American Statistical Association (ASA) published a statement about $p$-values to help researchers with their correct use [1]. Three years later, in 2019, an editorial in *The American Statistician* strengthened the message to move the world beyond "$p < 0.05" [2]. Although they did not discourage the use of $p$-values, they stated the following about statistical significance: "based on our review of the articles in this special issue and the broader literature...it is time to stop using the term statistically significant entirely. Nor should variants such as *significantly different*, "$p < 0.05\)\), and "*nonsignificant* survive, whether expressed in words, by asterisks in a table, or in some other way' [2].

Criticism of some uses of hypothesis testing and $p$-values is not new [3]. One of the main reasons for these recommendations is the use of statistical significance as a true/false decision boundary to select variables [4–7]. In the biochemical sciences, it is frequent to (i) derive biological inferences only from the biomolecules with statistically significant changes, (ii) report only the biomolecules with statistically significant changes (selective reporting), and (iii) report selected $p$-values or other test statistics as inequalities or asterisks. These practices may bias the biological interpretation and limit the generation of robust scientific knowledge. Consequently, many statisticians and researchers advocate for retiring statistical significance [2,8].

Instead of abandoning statistical significance, many would prefer to decrease the cutoff [9–12], specifically from $p < 0.05$ to $p < 0.005$ [13–15]. This change would reduce the number of false discoveries and improve the reproducibility of research [14]. This change would be simple to implement because most researchers already use hypothesis testing and statistical significance [14]. Nevertheless, it has been counterargued that decreasing the cutoff would not solve the problems [6]. For example, it has been highlighted that decreasing the cutoff for statistical

---

**Highlights**

- The use of statistical significance is under discussion. Many statisticians and researchers advocate for its retirement. Conversely, other statisticians and researchers think that its retirement would damage science.
- There is room for improvement in the use of hypothesis testing and $p$-values in biochemical sciences and omics.
- The selection of variables by statistical significance with solid cutoffs drives and may bias the biological interpretation of biochemical data.
- To obtain robust knowledge by comparing studies, it is essential to report thoroughly all results (both quantitative and categorical variables).
- Because of the big number of variables, the problems of selecting variables by statistical significance increase for omic studies.

---

1. Unidad de Excelencia, Instituto de Biología y Genética Molecular (IBGM), Universidad de Valladolid, Consejo Superior de Investigaciones Científicas (CSIC), Valladolid, Spain
2. Analytical Pharmaceutical Chemistry, Department of Medicinal Chemistry, Uppsala University, Sweden
3. Current address: IBGM, Calle Sanz y Forés 3, 47003, Valladolid, Spain

*Correspondence: david.balgoma@ilk.uu.se (D. Balgoma).
significance would not solve the selection of overestimated effect sizes (discussed in the next section) [16].

Some authors even indicate that abandoning statistical significance may damage science [17]. They consider that abandoning statistical significance would favor unjustified claims and mislead science [18]. Conversely, those that defend retiring statistical significance highlight that setting cutoffs for \( p \)-values has favored non-ethical practices (e.g., \( p \)-hacking) and damaged science. Although hypothesis testing does not lead to bad practices per se [18], current publication processes favor the publication of statistically significant results. In combination with current career incentives (e.g., ‘publish or perish’), statistical significance may favor conscious or unconscious bad practices in research.

We see that statisticians and philosophers of science have an ongoing debate about the role of statistical significance in statistical inference [2,14,16,18,19]. Beyond the controversies, it seems clear that all statisticians encourage a better use of hypothesis tests and \( p \)-values by researchers. These tools are powerful to improve research, but only if we understand them. To achieve better statistical inference and reproducibility in biochemical sciences, researchers must understand the logic behind hypothesis testing and statistical significance, including their limitations regarding our expectations. For example, many scientists believe that low \( p \)-values offer a high degree of certainty. However, \( p \)-values depend on your specific data (even in big and well-designed studies). Consequently, a replication study may generate a very different \( p \)-value. Only repeated experimentation confirms the discovery suggested by a very low \( p \)-value. Other misinterpretations of \( p \)-values, confidence intervals, and statistical significance are given in Box 1.

In the following sections, by means of simplified examples in the context of biochemical sciences and omics, we briefly introduce (i) how selecting variables by statistical significance affects the biological interpretation of data, (ii) how selective reporting damages the reproducibility of science and the generation of robust knowledge, (iii) the specificities of omics, and (iv) two alternatives to hypothesis testing and \( p \)-values. The interested reader may find a deeper discussion in the references that we provide.

The selection of variables by statistical significance drives and may bias the biological interpretation of data

Selecting variables by statistical significance focuses the inference on a subset of the ensemble of evidence, which may bias the biological interpretation of the dataset [20]. To illustrate this point in biochemical sciences, let us consider a hypothetical example of a study that measured fatty acids in plasma (Figure 1A). In this example (i) the fatty acids were quantified in a relative way, and (ii) the change was measured by the logarithm of the fold change between a treated group and a control group (i.e., below 0 means decrease, above 0 means increase). According to previous research, the increase in the levels of free fatty acids in plasma corresponds to an increase in lipolysis in the adipose tissue [21–23].

In this context, let us consider two different strategies to extract the biological interpretation from the simulated data in Figure 1A.

Strategy 1: selection of biomolecules by statistical significance when \( p < 0.05 \) (Figure 1A, left). Only fatty acids FA(20:4) and FA(20:5) were selected to derive a biological interpretation. These two fatty acids are polyunsaturated fatty acids (PUFAs) and their levels are increased. In the context of the study, we would deduce a strong selectivity for PUFAs during lipolysis in adipose tissue.
In addition, using strategy 1 and a more stringent cutoff for strategy 1, we would overestimate the selectivity of lipolysis of lipids with PUFAs in adipose tissue. Likely to be selected by statistical significance (Figure 1A, right). We observe that the average of the logarithm of the fold changes is higher than 0 in all cases (i.e., levels are increased). Considering the scientific context, we would conclude that this common behavior of fatty acids suggests an increase in lipolysis in adipose tissue. In addition, it seems that PUFAs experienced a stronger increase than did the other fatty acids. This behavior would suggest a certain selectivity for PUFAs, which would be confirmed or discarded by further experimentation.

We can see that the two strategies yield different biological interpretations. They also differ in how to address further research. Many statisticians recommend strategy 2, which focuses on the size of effects (e.g., fold change) of all studies rather than only on the statistically significant effects [3]. One of the main reasons is that statistical significance in isolated studies selects variables with inflated effect sizes [24]. Inflated effects arise because sampling variation may lead isolated studies to present a stronger effect than the true effect. In isolated studies, those inflated effects are more likely to be selected by statistical significance [24]. Consequently, selecting variables by statistical significance overestimates the scientific importance of the selected variables. In our example, with strategy 1, we would overestimate the selectivity of lipolysis of lipids with PUFAs in adipose tissue.

In addition, using strategy 1 and a more stringent cutoff for p-values such as p < 0.005, we would conclude that no fatty acid was changed. We would need to increase the sample size to better determine the fold change of the fatty acids. However, this is not usually possible. For example, it is not possible when studying a rare disease because of patient availability. In

**Box 1. An overview of the use and limitations of p-values, confidence intervals, and statistical significance**

What these statistical tools are and their correct use have been explained in detail before [1,2,4-6,26,49]. We refer to the texts of Emmert-Streib and Dehmer [50], and Amrhein et al. [51], for a deep description. In the light of the literature, we would like to highlight the following points.

**P-values**

In contrast to what many researchers understand, ‘p-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone’ [1,26]. Clear graphical explanations of p-values can be found at Amrhein et al. [52] and Goodman [4]. When the explanation is read carefully (see Glossary), one notes that a p-value refers to a probability of the test statistic assuming the null hypothesis.

**Confidence intervals**

The definition highlights that confidence intervals [3,52,53] should be interpreted as a long-run frequency description of samples from a population [43]. Similarly to p-values, confidence intervals entail the assumptions of frequentist statistics that considers probability as frequencies in large-scale resampling of a population (sampling distribution) [43]. Consequently, in contrast to what many researchers believe, we should not infer that the confidence interval from an isolated experiment has a 95% chance of containing the true value [26]. In addition, we should not interpret confidence intervals as intervals that will contain the parameter with a given probability (i.e., Bayesian credible intervals) [43].

**Statistical significance**

A variable that presents a p-value (or other test statistic) below a predefined threshold is said to be statistically significant (e.g., a variable with ρ = 0.0049 when the predefined threshold is ρ < 0.055). Statistical significance was introduced as a tool to suggest interesting results and perform further confirmatory research [54]. However, statistical significance is often used as a true-difference/no-difference decision boundary. Subsequently, the variables selected as being truly changed are considered to be scientifically significant. In opposition to this practice, statisticians highlight that p-values or equivalent test statistics should not be interpreted as arbitrarily thresholded values [55].

From this overview, we see that p-values, confidence intervals, and statistical significance are limited in comparison with the expectations of many researchers. As recently and clearly summarized by Berner and Amrhein [24], (i) p-values depend on your data and are not reliable, (ii) an absence of statistical significance does not indicate an absence of an effect, (iii) statistical significance does not select reliable effect sizes (e.g., fold changes), and (iv) statistical significance does not equal scientific significance.

**Storey’s q-values**: a tool to control false discoveries when multiple hypothesis tests are performed (as in omics). If variables are independent and the null hypothesis is true (no difference in a t-test), p-values present a uniform distribution between 0 and 1 [67]. Variables with a true difference tend to present low p-values and cluster close to 0 in the histogram. This characteristic allows the rate of false discoveries to be controlled [67].

**Test statistic**: this summarizes the data and quantifies the distance between the data and the model assumptions and predictions [26]. In the most common t-test in biochemical sciences, a t-statistic summarizes the data by first computing the difference between the means of two groups, and second by scaling this difference by the standard error. Consequently, the t-statistic quantifies the distance of the data to the assumptions of the test, including that there is no difference between the means of the two groups (null hypothesis).
a common case in research, it is not feasible to increase the number of animals because of ethical considerations.

In conclusion, rather than thresholded test statistics (e.g., \( p < 0.05 \)), thoughtfulness and context are necessary to extract the biological information [25]. In our example, one should be very thoughtful before concluding enzyme selectivity by using statistical inference in an isolated study. Only repeated experimentation and mechanistic research can confirm it.

Naturally, \( p \)-values can still be used ‘as a continuous measure of the compatibility between the data and the entire model used to compute it’ [26]. By entire model, Greenland et al. mean the assumptions of the hypothesis test, which include the null hypothesis. The most common case in biochemical sciences corresponds to a \( t \)-test applied to the fold change between two groups: the null hypothesis assumes no difference between the two groups. In this general case, a low \( p \)-value can be seen as a low compatibility of the data with no difference (given the rest of assumptions). Consequently, low \( p \)-values do not provide certainty, and instead provide a continuous (non-dichotomous) degree of compatibility with no difference of the biomolecule between the two groups under study. Because of the lack of certainty of \( p \)-values, mechanical application of statistical significance to select biomolecules should be avoided.

**The more the better: complete reporting to improve reproducibility**

Another practice associated with statistical significance is selective reporting: only the results that are statistically significant are communicated [7]. This practice limits comparison among studies and induces a report/publication bias [15]. To illustrate this point, let us analyze a simplified
example of two simulated studies that measure two biomolecules in the same biological process (Figure 1B). In both studies, the fold changes were tested against no difference by t-tests. When both studies perform selective reporting with \( p < 0.05 \) (Figure 1B, left), a comparison between the two studies is impossible: study 1 reports only biomolecule A; study 2 reports only biomolecule B.

It seems they do not reproduce each other. By contrast, if both studies perform complete reporting (Figure 1B, right), the effects of both studies agree and seem to be reproducible.

From this example, we see that complete reporting of data enables the integration of different studies and the generation of robust knowledge (Box 2). To integrate data from different studies, we should consider that other researchers need our raw data, not only a summary in a graphical way or as the average ± standard deviation. For every observation, quantitative variables should be reported as numbers and categorical variables as their possible values. Replication data are nested and can be integrated by, for example, multilevel models [27]. Although not all biochemistry practitioners need to be experts in integrating data, all of us should understand the importance of sharing our data such that our research does not wane. For example, it is very difficult to integrate new data with the data from 50-year-old biochemical studies. It is very rare to report data for all observations and the corresponding authors of old publications are no longer available.

From the example in Figure 1B, we also see that complete reporting of all statistical analyses characterizes each statistical inference in a complete way. From this point of view, if one uses hypothesis tests, it is strongly recommended to report all tests performed, not only those that were statistically significant. Consequently, we should report (i) all test statistics, (ii) their associated degrees of freedom, and (iii) all \( p \)-values. Regarding the format of reporting statistical analyses, test statistics and \( p \)-values should be reported as exact values (not as inequalities or asterisks; e.g., \( p < 0.005 \) or ***) [2].

In a real example of the importance of complete reporting, we observed that a specific group of lipids was consistently upregulated in different studies analyzing human plasma from patients with different diseases [28]. All these diseases display upregulated de novo lipogenesis in the liver. The different authors reported upregulation of these lipids, even if their changes were not statistically significant. Consequently, the openness of the authors allowed us to perform inductive reasoning and identify these biomarkers [28].
In conclusion, complete reporting of data allows researchers to compare and integrate their studies. This is fundamental to establish a consensus and generate robust knowledge (Box 2) [29].

**Growing pains in omics**

Omic studies measure a large number of biomolecules (RNA transcripts, proteins, polar metabolites, and/or lipids). In addition, when the measurements are done by mass spectrometry, it is relatively frequent to analyze features (unidentified combinations of retention time and m/z). Omic studies perform a comprehensive characterization of biological systems, but the large number of variables (i) increases the probability of detecting false differences (false positives), and (ii) aggravates the effects of selecting variables by statistical significance.

**False discovery rate (FDR)** procedures aim to control the ratio of false positives when multiple hypothesis tests are performed. *Storey’s q-values* are very frequently used in omics, but this tool requires independent variables [30]. However, the biomolecules in omic datasets are not independent. For example, in a cell model, the lipids PC(sn-16:0/20:4) and FA(20:4) are directly related by, at least, the activity of phospholipases A2 [31,32]. In our experience, this intrinsic dependence in the omic data usually leads to ‘U-shaped’ p-value histograms. This indicates that the requirements of Storey’s q-values are not met and they have a limited applicability in omics. Consequently, researchers should consider FDR variants that consider dependency [33–35]. In any case, no FDR test statistic should be used in association with true/false decision boundaries to select variables. As described for p-values before, the selection of variables by statistical significance based on q-values or other FDR statistics would drive and potentially bias the extraction of the biological information contained in the omic data.

Despite these considerations, because of the myriad of variables in omics, it may be impossible to consider the complete omic dataset to derive a biological interpretation. As the first strategy for variable reduction, we could estimate how different decision boundaries affect the inference. For example, the biomolecules could be ranked by their fold change and evaluate whether we would infer the same biological interpretation when \( p < 0.005, p < 0.05, p < 0.1, \) and \( p < 0.25 \) (these cutoffs should be considered as a mere suggestion). By using different cutoffs, no strong decision boundaries are considered and the number of variables is reduced.

Another strategy for variable reduction is the application of principal component analysis (PCA) [36]. PCA yields a set of values (scores) of new, orthogonal, variables that are combinations (loadings) of the original variables (biomolecule signals). The new variables (i) are ranked by the variation that they explain, and (ii) are orthogonal (i.e., they are statistically independent) [36]. To interpret the principal components (i) the representation of the loadings may allow us to assign the different principal components to different biochemical pathways, and (ii) the representation of the scores against the treatments may allow us to estimate the regulation of the pathways that the principal components represent. In our experience, principal components usually cluster biomolecules produced by the same pathway, which simplifies the interpretation (e.g., Balgoma et al. [21,37]). Because principal components are orthogonal (statistically independent), Storey’s q-values could be applied to the scores. Nevertheless, no strong decision boundaries should be applied to the test statistics to select which principal components are scientifically significant. Finally, PCA should not be applied in a mechanical way: ‘PCA results may not be as reliable, robust, or replicable as the field assumes’ [38]. Similarly to any statistical tool, consider that the results of PCA depend on the specific dataset of an isolated study.

Finally, comparisons of omic studies also present ‘growing pains’ when selective reporting by statistical significance is performed. Let us consider two well-designed omic studies about the
same disease. Let us imagine that both studies (i) measure the same biomolecules, (ii) present the same number of replicates, and (iii) use the same cutoff for statistical significance to report biomolecules. One would expect the two studies to reproduce each other. However, \( p \)-values depend on the estimated effects and the estimated variation in the samples of a specific study. Consequently, \( p \)-values are ‘fickle’ when calculated from different studies and are not reliable [24,39,40]. Because a large number of biomolecules are analyzed in omics, these two identical omic studies may report very different datasets of biomolecules if they perform selective reporting by statistical significance. Similarly to the case in Figure 1B for two variables, the two studies may seem not to reproduce each other, when in fact they do.

In conclusion, omics practitioners should be aware that the problems associated with the selection of variables by statistical significance and selective reporting are of special importance when we deal with a myriad of variables. There is no perfect recipe to solve the challenges of omics. Nevertheless, PCA or the application of statistical significance to select variables should not be done in a mechanical way. Similarly to the general case, context and comparison with previous knowledge should prevail in omics.

Beyond \( p \)-values and statistical significance: size of effects and Bayesian framework

Different tools to help researchers in data analysis have appeared over recent centuries (Box 3). Although null hypothesis tests and \( p \)-values (or other test statistics) are the most common tools for data analysis, other perspectives are possible and legitimate [41]. For example, we analyzed our data considering fold changes and their confidence intervals without using \( p \)- or \( q \)-values [21,42]. This corresponds to strategy 2 in Figure 1A.

The Bayesian framework offers another option for statistical analysis. A detailed description of Bayesian statistics is beyond the scope of this text and we refer the reader to the textbook of McElreath [43]. Briefly, in the general context of biochemical sciences, the parameter of interest is the fold change of a biomolecule. According to our prior knowledge, we assign a prior distribution of values to the fold change. For example, because we do not expect a strong difference in

**Box 3. A brief historical vision of statistical inference**

Statistical inference is a field with two main frameworks: frequentist and Bayesian statistics [46]. Historically, there were Bayesian applications in early science but advances in frequentist statistics led this framework to dominate in the 20th century [57].

In addition, frequentist statistics developed two main schools of thinking about statistical inference: (i) the school founded by Fisher, and (ii) the school founded by Neyman and Pearson [51,57]. Although the details are beyond the scope of this manuscript, these two schools of thinking have incompatible foundations and interpretations. However, the so-called ‘statistical rituals’ combine elements of these two opposing frequentist practices [57–60]. Despite the differences, the two schools agree about (i) using statistics to help researchers to suggest interesting results in the context of their discipline, and (ii) performing repeated experimentation to confirm the discovery [54].

Consequently, hypothesis testing and \( p \)-values were not introduced to decide which variables are scientifically relevant [54]. However, it is relatively common to use statistical significance as an equivalent to scientific significance [57–60]. This use was already controversial more than a century ago [61]. In combination with perverse career incentives and non-ethical practices (e.g., \( p \)-hacking), selecting variables by statistical significance is part of the crisis of reproducibility of science in the 21st century [15].

Although Bayesian methods were despised by Fisher as ‘founded upon an error, and must be wholly rejected’ [52], Bayesian statistics is another powerful tool for researchers. The differences between Bayesian and frequentist statistics are subtle and a full description is beyond the scope of this manuscript. We refer the reader to previous publications for a deep explanation [45,63,64]. In recent years Bayesian statistics has become more popular. Nevertheless, Bayesian statistics is highly demanding from a computational point of view. This drawback is more pronounced in omics because of the large number of variables.
membrane lipids or the transcript of a housekeeping gene, we would assign a prior distribution that is skeptical about differences of an order of magnitude. Importantly, prior distributions should not be 'tweaked' according to the data (a bad practice similar to p-hacking). Finally, the data update these prior distributions to obtain the posterior distributions of the fold changes.

One of the main advantages of the Bayesian framework is that the posterior distribution of the parameters and their associated credible intervals correspond to the distributions of the plausibility of the values of the parameter [44]. In fact, many researchers wrongly interpret frequentist confidence intervals as Bayesian credible intervals. From this point of view, Bayesian statistics fits the expectations of most researchers. Nevertheless, one should consider that when the number of samples is low, prior distributions may dominate the results.

Finally, in the context of Bayesian modeling, complete reporting of data analyses requires communicating (i) the priors used for the fold changes, and (ii) the posterior distribution of the fold changes (or a summary, such as highest posterior density intervals [43,45]).

Concluding remarks

We would like to highlight that we do not recommend a specific strategy to analyze biochemical or omic data. Authors may use hypothesis tests and p-values. All alternatives have pros and cons. Nevertheless, statistical tools should not be used in a mechanical way (e.g., selecting variables by statistical significance). Considering the controversies, several unresolved issues remain about the future relationship between statistical inference in biochemical research and publication policies (see Outstanding questions).

To conclude, we have the following take-home messages to improve data analysis in biochemical sciences and omics. (i) Statistical inference is not a fixed discipline and evolves in time. To improve the extraction of the biological information from your biochemical and omic datasets, follow the discussions about scientific inference by statistical analysis [2,18,20,46–48]. (ii) Avoid selecting variables by statistical significance on solid cutoffs to derive an interpretation of your biochemical data. Interpret all your biochemical data in the context of previous knowledge and the known relationships among biomolecules [25]. (iii) To improve reproducibility and enhance robust science, perform complete reporting of your data (quantitative variables as numbers, categorical variables as their possible values). (iv) If you use hypothesis tests, report all your test statistics and p-values as exact values [2].

Acknowledgments

We kindly thank Dr Marta Montero Crespo and Andrés Martinez Arroyo for insightful comments. D.B. acknowledges support from a María Zambrano Fellowship (Next Generation EU), the Proyecto de Internacionalización de la Unidad de Excelencia IBGM (CL-ES-2021 IBGM), and the Programa Estratégico IBGM (CCVC8485). This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of interests

No interests are declared.

References

2. Wasserstein, R.L. et al. (2019) Moving to a world beyond ’p < 0.05’. Am. Stat. 73, 1–19

Trends in Biochemical Sciences

Outstanding questions

Will researchers implement the improvements in the use of hypothesis tests and p-values that statisticians strongly suggest?

Will scientists abandon statistical significance?

Will biochemical and omics journals retire or reduce the use of p-values and/or statistical significance?

Will journals and editors in biochemical sciences require reporting all test statistics and p-values as continuous quantities?

Would retiring statistical significance be more likely to damage science or improve it?

Will biochemists share all their data in a thorough way?

Will Bayesian analyses become more popular?
34. Balgoma, D. et al. (2022) Orthogonality in principal component analysis allows the discovery of lipids in the jejunum that are independent of ad libitum feeding. Metabolites 12, 866
35. Elshah, E. (2022) Principal component analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated. Sci. Rep. 12, 14663
41. James, G. et al. (2013) An Introduction to Statistical Learning: With Applications in R, Springer
42. Balgoma, D. et al. (2022) Orthogonality in principal component analysis allows the discovery of lipids in the jejunum that are independent of ad libitum feeding. Metabolites 12, 866
55. Balgoma, D. et al. (2022) Orthogonality in principal component analysis allows the discovery of lipids in the jejunum that are independent of ad libitum feeding. Metabolites 12, 866
56. Balgoma, D. et al. (2022) Orthogonality in principal component analysis allows the discovery of lipids in the jejunum that are independent of ad libitum feeding. Metabolites 12, 866