PROF. HARALD MISCHAK (Orcid ID : 0000-0003-0323-0306)

Article type : Editorial

Datasharing: obsolete? impossible in times of GDPR ? Or mandatory in science?!

Scientific research and progress is based on the principle of hypotheses supported or disproved by repeated testing and scientific prove, which in turn is based on data. This principle has been implemented also in the context of scientific publishing. In the past, we have all adhered to the principle that within the results' section of a manuscript we present the prove for our hypothesis and scientific evidence. However, as a result of the increasing amount of data generated and consequently used to support the conclusions, it became difficult and subsequently impossible to present all underlying data in a manuscript. Owed to the requirement to demonstrate correctness and validity of the published work and the conclusion drawn from the actual data, and supported by the development of on-line publishing, the publishing of supplementary material and depositing the actual data in data repositories became increasingly popular. However, at the same time, and also due to the frequently large size of datasets, accessibility to the actual data was not regarded a prerequisite, and the requirement to give access to the data, to prove correctness and validity of the claims raised in a publication, was increasingly neglected. As also depicted in **Figure 1**, especially in the era of "omics" huge amounts of raw, machine data are being generated, resulting in file sizes exceeding, by far, 1 Gigabyte. As these datasets cannot be evaluated manually, even more as thousands of such datasets are being generated for a complete study, consequently the actual data for one single publication reaching multiple Terabytes. Consequently, software solutions are employed for data interpretation, to reduce the

huge amount of data and condense the information to e.g. a list of compounds and their individual abundance. These datasets are ultimately further condensed into e.g. a list and combinations of features, or even single features being reported as significantly changed under certain conditions, for example when comparing specific diseases to controls. The latter is in general reported in a manuscript as the result of the experimental approach. This approach is certainly practical, but at the same time contains multiple options for error or misinterpretation, in the worst case even falsification. This danger of ultimately reporting incorrect results is even more pronounced once the actual raw data are not being disclosed. In such a case, the readers and the scientific community can do nothing else than accept the interpretation of the data by the scientists reporting them as absolutely correct. Reproducing and verifying these results is impossible, as the actual underlying data are not available.

When critically evaluating the flow of information in a study as depicted in Figure 1, it becomes obvious that we rely on the correctness and accuracy of the applied software solutions, especially for the first step, when the information is extracted and condensed from the raw machine data. However, this assumption is not necessarily correct, as different software solutions produce different output of the same data input, therefore eventually leading to different conclusions and results. The risk of misinterpretation is even further enhanced as a result of the frequent application of the option to adjust multiple parameters (e.g. signal/noise or background subtraction, accepted levels of confidence, etc.). In addition, in the next step, different types of statistical assessment, of adjustment for multiple testing, of exclusion criteria, etc., can be applied, again of potentially enormous impact on the results obtained. In conclusion, the same raw data may well ultimately be condensed into very different results, depending on the solutions applied. This is even further enhanced by the pressure on scientists to publish, introducing a bias towards identifying positive, significant result (we all are aware of the fact that a study reporting no significant results is in general difficult to publish).

All the above facts and issues contribute to the reluctance towards data sharing that is a prominent problem ever since, as in general nobody is tempted to invite severe critique. In addition, data sharing also takes away ownership of the data from the individual scientist, which also does not raise the interest in data sharing, even if this practice in fact is ethical, especially if data are generated using public or social security money.

This situation has recently substantially worsened, as a result of the introduction of the general data protection rule (GDPR) in Europe, which is echoed by similar yet not as detrimental legal frameworks in other countries. The introduction of GDPR, while theoretically likely starting from the positive aim to protect individuals from exploitation, by now has had a severe negative impact on science (and likely also on other areas, not the topic of this article) [1]. GDPR has been (ab)used to refuse sharing of raw data. As a result, interpretation of the actual data collected in a study is left solely to the scientists conducting the study, any re-evaluation or attempt to reproduce results by the scientific community is not possible in such a case, e.g. [2,3]. Additional negative side-effects are that the data, even though generated with public funding, are not accessible for the public who actually paid for the generation of these data in the first place. The consequence is also that similar data have to be generated again, in case they are required for any further experiment. Such an approach is inappropriate, and should neither be tolerated by (public) funders of studies, nor be supported by publishers.

A recent and good example of the seriousness of this problem, and also of an appropriate way how to move towards resolving this issue is a manuscript by Gatenholm et al on the identification of peptides associated with osteoarthritis [4]. The authors aimed at identification of potential peptide biomarkers to better understand the development of osteoarthritis and pain. Based on the data generated, the identification of 6292 endogenous peptides was reported, with 566 peptides differing significantly in unwounded compared to wounded osteoarthritic zones. Based on these results, it was concluded that with further refinement, peptidomics could potentially become a diagnostic tool for osteoarthritis, and improve the knowledge of disease progression and genesis of pain. Upon request and after ascertaining that the data can be shared even in times of GDPR, the authors made the raw data available for evaluation with a different, also widely used software solution for mass spectrometry data assessment. Surprisingly, after repeating the analysis using a different software solution even when applying identical search parameters, the results were very different from the results reported by the authors in the published manuscript [5]. It is currently unclear which of the two reported results are more accurate (obviously absolute correctness is an ideal that rarely if ever can be reached). However, on this occasion the correct approach by the authors (from the view of scientific ethics) enabled spotting errors in at least one of the used software packages, and possible correction of any

conclusion drawn. This would not have been possible if the raw data were not shared. Further, this example very well demonstrates the huge danger associated with avoiding sharing of data and reporting only conclusions and results that may be based on flawed software packages (both software packages used in this case are commercial products and widely used, however, at least one must contain major flaws).

This example supports the notion that publishers of respectable scientific journals should implement a strict requirement to enable access to all underlying data. Publishing of "results" without giving full access to all data that were used to generate the results should not be acceptable, and has the potential to ultimately cause huge harm to science, by putting into question the credibility of data. Of note: it may not always be easy to deposit data in repositories, but this does not prevent sharing the data upon request, e.g. by ftp or even via shipping of hard discs. An excellent positive example that data sharing is in fact possible is The Cancer Genome Atlas [6], but multiple other examples exist as well. Funders should pay special attention to this issue: public funding should not be used to generate private data. Such strict requirement may give a disadvantage (e.g. lower number of submission) to the journals implementing such rules in comparison to other journals aiming at publishing any article irrespective of their scientific merits (sometimes also referred to as predatory journals), but, on the other hand, it will ultimately substantially improve the quality, validity and credibility of science.

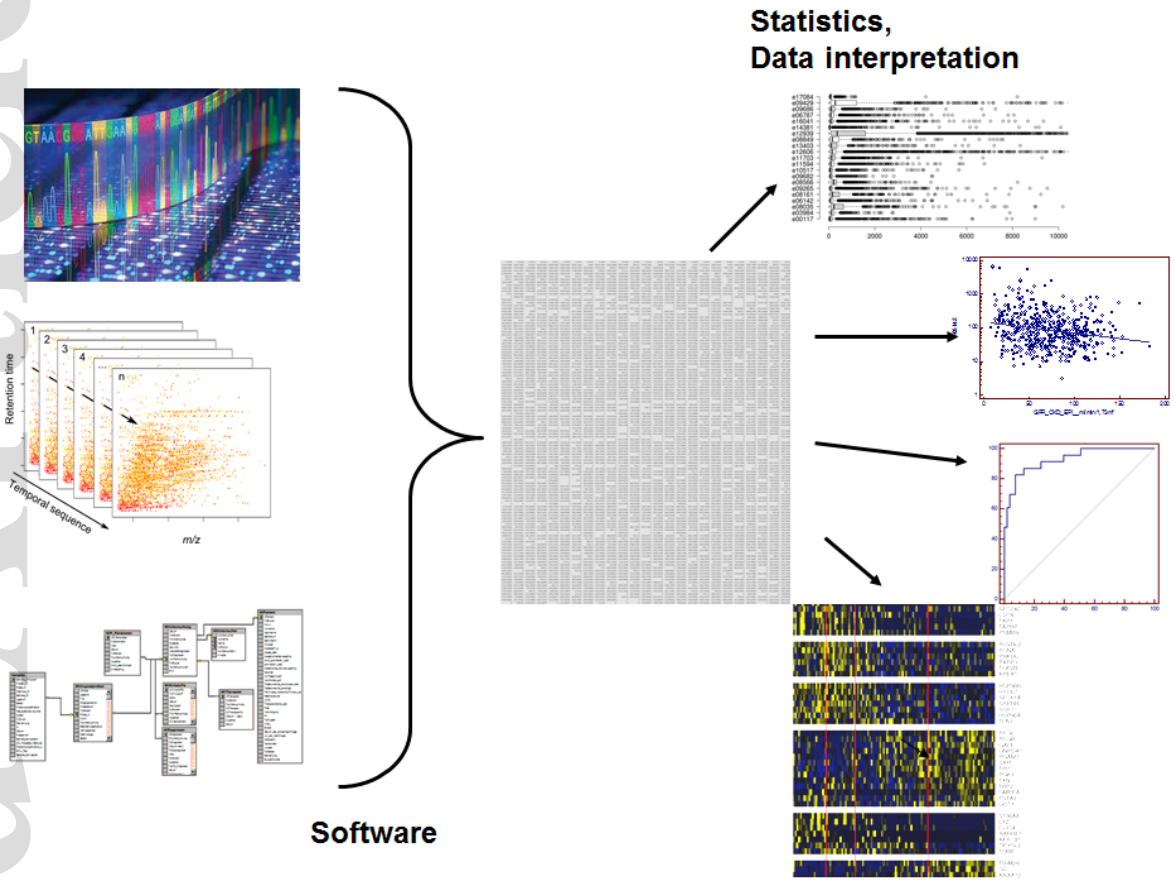

Figure 1: typical experimental flow in multiparametric (omics) experiments. TB of raw data (left) are evaluated based on specific criteria and using different software solutions, to generate a list of compounds. This list is further assessed using e.g. different statistical methods and interpreted, to ultimately be condensed into a few reported results.

REFERENCES

[1] E. Critselis. *Proteomics Clin Appl.* 2019, **13**, e1800199.

[2] A. Christensson, J.A. Ash, R.K. DeLisle, F.W. Gaspar, R. Ostroff, A. Grubb, V. Lindström, L. Bruun, S.A. Williams. *Proteomics Clin Appl.* 2018, **12**, e1700067.

[3] S. Ravizza, T. Huschto, A. Adamov, L. Bohm, A. Busser, F.F. Flother, R. Hinzmann, H. Konig, S.M. McAhren, D.H. Robertson, T. Schleyer, B. Schneidinger, W. Petrich. *Nat. Med.* 2019, **25**, 57.

[4] B. Gatenholm, J. Gobom, T. Skillback, K. Blennow, H. Zetterberg, M. Brittberg. *Eur. J. Clin Invest* 2019, **49**, e13082.

[5] M. Pejchinovski, Mischak H. *Eur. J Clin Invest* 2020, **in press**.

[6] J.N. Weinstein, E.A. Collisson, G.B. Mills, K.R. Shaw, B.A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J.M. Stuart. *Nat. Genet.* 2013, **45**, 1113.