BIOINFORMATICS LETTER TO THE EDITOR Vol. 26 no. 3 2010, pages 437–439
doi:10.1093/bioinformatics/btp648

Gene expression

Over-optimism in bioinformatics research

Anne-Laure Boulesteix

Department of Medical Informatics, Biometry and Epidemiology, University of Munich, Marchioninistr. 15, 81377 Munich, Germany

Received on October 23, 2009; revised on November 10, 2009; accepted on November 11, 2009

Advance Access publication November 26, 2009

Associate Editor: Martin Bishop

Contact: boulesteix@ibe.med.uni-muenchen.de

The problem of 'false research findings' in medical research has focused much attention in the last few years (Ioannidis, 2005). One of the main problems, termed as 'fishing for significance' in the present letter, is that researchers often (consciously or subconsciously) report results that are in fact the product of an intensive optimization, i.e. of multiple comparisons. Such results are typically unlikely to be reproduced in an independent study and have a high probability to be false (Ioannidis, 2005). The 'fishing for significance' problem is enhanced by the so-called 'publication bias': positive results have a much higher chance to get published than negative results, as already acknowledged 50 years ago (Sterling, 1959).

In a word, many false positive results are produced through multiple comparisons, and false positives have higher chance to get published than true negatives. Moreover, the difficulty to publish negative results obviously encourages authors to find something positive in their study by performing numerous analyses until one of them yields positive results by chance, i.e. to fish for significance. Although this issue is by far less acknowledged and publicly admitted than in the medical context, the same types of problems occur in biostatistics and bioinformatics research.

1 FISHING FOR SIGNIFICANCE

In a recent editorial of the journal *Bioinformatics* on 'Papers on normalization, variable selection, classification or clustering of microarray data', Rocke *et al*. (2009) state that 'prediction methods enter a crowded area'. Indeed, hundreds of prediction algorithms for high-dimensional small-sample data have been proposed in statistics, bioinformatics and machine learning journals in the last few years. Rocke *et al*. (2009) further claim that 'consciously or subconsciously, the developer of a new method optimizes its characteristics against the datasets to be used for evaluation'. This is because the development of a new prediction algorithm is often a trial-and-error learning task. The emerging method is successively adapted depending of the intermediate results. This problem can be paralleled to the 'fishing for significance' issue in medical research, except that in bioinformatics research the researcher fishes for an improvement (e.g. a decrease of the error rate) instead of fishing for a significant *P*-value.

In statistical bioinformatics research, fishing for significance consists of two distinct components: (i) the sequential adaptation of the new methods to the considered datasets as described in the Bioinformatics editorial, and (ii) the search for a specific dataset or simulation setting for which the new method works better than existing approaches as quantitatively investigated in the recent paper by Yousefi *et al*. (2009). Both mechanisms lead to optimistic conclusions regarding the superiority of the new method.

The first mechanism essentially affects all research fields related to data analysis such as statistics, machine learning or bioinformatics. The trial-and-error process is indeed an important component of data analysis research. One would not expect a statistician or bioinformatician to develop a method with pen and paper, try it once on a dataset and immediately write a paper. Most original good ideas have to be sequentially improved before reaching an acceptable maturity. The development of a new method is *per se* an unpredictable search process. Thus, the concept of analysis plan known from medical and pharmaceutical research cannot be easily accommodated to bioinformatics research. The problem is that, as stated by the *Bioinformatics* editorial team, this search process leads to an artificial optimization of the method's characteristics to the considered datasets. Hence, the superiority of the novel method over an existing method (as measured, e.g. through the difference between the cross-validation or bootstrap error rates) is sometimes considerably overestimated. This problem potentially occurs in all data analysis problems with a clearly defined and objective quality criteria.

In a concrete medical prediction study, fitting a prediction model and estimating its error rate using the same training dataset yields a downwardly biased error estimate commonly termed as 'apparent error'. Validation on independent fresh data is an important component of all prediction studies. Similarly, developing a new algorithm and evaluating it by comparison to existing methods using the same datasets may lead to optimistically biased results in the sense that the new algorithm's characteristics overfit the used datasets. The over-optimistic result is the superiority of the new algorithm compared with existing methods (for instance, in terms of prediction error) rather than (like in concrete prediction studies) the prediction error itself. In the same way as a prediction rule has to be validated using fresh data in applied research, one can try to validate the superiority of the new algorithm in methodological research. While this idea may appear at first glance as a trivial generalization of the validation concept from medical research, it raises its own methodological difficulties. Getting to relevant validation datasets is generally not a problem, in contrast to what

happens in applied biomedical research. A plethora of data of all types can be found in the WWW on public repositories, journal web sites or homepages of the researchers. Completely new data types are an important exception. But in most cases, the main problem is rather the definition of eligibility criteria than the search for datasets. Authors may be consciously or subconsciously tempted to select a particular dataset because it is excepted to yield better results with the new method.

This leads us to the second component of the fishing for significance mechanism: the biased selection of datasets that makes the new algorithm look better than it actually is. This biased selection may occur at different levels. For example, authors may deliberately omit to report the results obtained with a particular dataset just because existing methods outperform the new algorithm on these data. This kind of reporting bias is quantitatively investigated in the remarkable study by Yousefi *et al*. (2009) with striking results. Selecting the example datasets based on their results yields a substantial optimistic bias in error rate estimation. A less extreme variant of this scenario is when authors choose to ignore a dataset in their study because they suspect that, based on theoretical considerations or past experience from similar studies, their new algorithm will not perform well on this particular data structure. For instance, a prediction method for high-dimensional data may be expected to perform badly in the case of many very weak covariates or conversely in the case of few very strong covariates. In contrast with the deliberate omission of bad results, the omission of datasets that are expected to yield bad results may be correct as long as the authors state in their paper that the method is especially designed for a particular data situation but may be less appropriate in other cases. Conversely, it would be misleading to highlight the general character of the method but evaluate it only on a particular type of data which is expected to yield advantageous results. It is not wrong to focus on a particular data type, but these restrictions should be well documented and openly admitted. This problem is related to the theoretical assumptions underlying a method as extensively discussed by Mehta *et al*. (2004). It is unreasonable to expect that a new method performs universally better than existing methods in all settings: most methods perform well under certain assumptions only. It is thus important to specify the assumptions of a new method including both theoretically motivated assumptions and restrictions that were identified empirically. That said, evaluating the effect of violations may also be interesting, especially in the case of very restrictive assumptions (Mehta *et al*., 2004).

To sum up, two mechanisms combine to yield over-optimistic results. First, the new method's characteristics often overfit the datasets which were used for its development. Secondly, when a new method is evaluated after its development, the biased selection of datasets also leads to over-optimistic results. The first problem is essentially inevitable. With respect to the second problem, the realistic definition and reporting of the area of application of the new method and the systematic selection of test datasets within this area should be given much attention in order to avoid over-optimistic results. Most importantly, we outline the importance of the two-stage approach: first, develop the method using example datasets and define its area of application as precisely as possible, then evaluate the developed method based on other datasets within the area of application (and report all the obtained results). This workflow, which is already consciously or subconsciously adopted by some authors, should perhaps be applied more stringently and consistently to ensure proper validation of research results.

The difficulty to unbiasedly select eligible datasets and describe the field of application of the new method is certainly enhanced by the fact that journals accept mostly positive research results (a method that performs better), sometimes neutral results (a review or a comparison of existing methods), but almost never negative results (a promising and sensible method that finally does not fulfill the authors' expectation), which leads us to the second major cause of false published research findings: the publication bias.

2 PUBLICATION BIAS

Publication biases and the necessity to 'accentuate the negative' (PLoS Medicine Editors, 2009) are well-documented in the context of medical and pharmaceutical research. Much effort has been taken to reduce the publication bias, for instance the prepublication registration of trials—often without much success (Ross *et al*., 2009).

Similarly, publication biases also affect studies on new bioinformatics or statistical methods, perhaps even more drastically than medical studies. A well-designed medical study with negative results has a reasonable chance to get published, at least in a low-impact journal. In contrast, a study on a new statistical method that turns out to be worse than existing methods in terms of objective performance will be rejected by almost all journals. In practice, the only way to present such negative results to the community is to include them in a comparison study.

As an example, let us consider a novel method A that 10 independent researcher teams consider as promising, but that is in fact not better than existing approaches. Eight of the 10 researcher teams correctly find that the promising method is finally not better. They forget it and do not write any paper. The ninth one finds a particular dataset on which method A performs better and this positive result gets published. The 10th one fishes for significance and publishes a variant A' that performs better than existing methods on his datasets. Note that in a subsequent validation based on an independent dataset, method A' would perhaps have not been better than existing methods, but here we assume that the authors did not perform such a strict validation. Finally, two papers documenting the superiority of method A/A' are published, although eight researcher teams found that it is not better than existing approaches. This example is probably caricatural, but similar things may occur in real life.

Of course, one could argue that, if method A is really bad, other researchers will not further use it and the community will finally find out that the method is bad. However, it would certainly have been better to publish the negative studies for different reasons. The ability of the method to establish itself also depends on various factors (such as the availability of software, the personality of the authors, etc.), such that it may take much time to find out that method A is actually bad. Moreover, as stated in the instructions for authors of the recently initiated *Journal of Interesting Negative Results in Natural Language Processing and Machine Learning*, 'knowing directions that lead to dead-ends in research can help others avoid replicating paths that take them nowhere' and 'much can be learned by analyzing why some ideas, while intuitive and plausible, do not work. […] Negative results may point to interesting and important open problems'. Hence, the publication of well-conducted studies

with negative result may be more useful than commonly assumed by journals and reviewers.

Moreover, publishing negative studies would potentially give more importance to qualitative aspects of new methods such as their conceptual simplicity, computational efficiency, interpretability, flexibility, ability to generalize or fit in a global framework, the absence of unplausible assumptions or, most importantly, the originality of the addressed research question. A negative aspect (an error rate that is slightly larger than those of existing methods) may be counterbalanced by positive aspects. For instance, in the context of prediction using high-dimensional data, a sparse method that can handle highly correlated variables may attract users even if its accuracy is not better (or even slightly worse) than existing methods. In the same vein, some studies that are negative in terms of the objective evaluation criterion (such as the error rate) may in fact contribute to the scientific progress as much as other studies with positive results because they suggest a completely new class of promising methods whose variants may eventually perform better in further studies.

Last but not least, partially relaxing the requirement for quantitative improvement (e.g. in terms of error rate or power) may in the long-term encourage honest and unbiased reporting and reduce the temptation to fish for significance. One could argue that it would also give more space to subjectivity in the review process. The decision whether a method with disappointing results was originally promising and whether readers may be interested in the negative conclusion may indeed be quite subjective. Assessing whether a new method with negative results is worth publishing is anything but trivial and suitable criteria should be carefully defined. But if negative results are systematically excluded from publication, authors are virtually urged to make their results seem positive and the reviewers' task is hindered by biased reporting, which also implies much subjectivity. Whether the referee should 'believe' the apparently positive results or not is a highly subjective question. Hence, the difficulty to objectively evaluate a negative study may be counterbalanced by an increased reporting transparency and a better application of appropriate validation procedures. On the whole, I believe that the occasional publication of well-designed studies on promising sensible ideas with disappointing quantitative results may in the long run contribute to a less optimistically biased literature. In this sense, the recently launched journals publishing negative results are an important step forward.

The publication of analysis scripts with the aim of research reproducibility (Hothorn *et al*., 2009; Peng, 2009) may also greatly contribute to more transparent reporting and enable a rapid validation of research results. Note that the availability of computer codes for reproducing the obtained results *does not* ensure that the authors did not overfit their method to the analyzed datasets. However, reproducible analyses considerably simplify the post-publication unbiased validation of the research findings. If the method can be quickly tested by running a well-documented script, readers can easily find out whether it performs as well as claimed in the article with the considered data types, or e.g. identify parameters that were tuned consciously or subconsciously. In this sense, a stringent reproducibility policy may, among many other advantages, reduce the temptation to fish for significance and help researchers to adopt a somewhat more objective point of view on their own studies. In this perspective, the publication of computer codes for the purpose of reproducibility should be expressly encouraged by editors and referees.

ACKNOWLEDGEMENT

I thank Carolin Strobl and Nicole Krämer for helpful comments.

Funding: LMU-innovativ Project BioMed-S: Analysis and Modelling of Complex Systems in Biology and Medicine.

Conflict of Interest: none declared.

REFERENCES

Hothorn,T. *et al*. (2009) Biometrical journal and reproducible research. *Biom. J.*, **51**, 553–555.

Ioannidis,J.P.A. (2005) Why most published research findings are false. *PLoS Med.*, **2**, e124.

Mehta,T. *et al*. (2004) Towards sound epistemological foundations of statistical methods for high-dimensional biology. *Nat. Genet.*, **36**, 943–947.

Peng,R. (2009) Reproducible research and biostatistics. *Biostatistics*, **10**, 405–408.

Rocke,D.M. *et al*. (2009) Papers on normalization, variable selection, classification or clustering of microarray data. *Bioinformatics*, **25**, 701–702.

Ross,J.S. *et al*. (2009) Trial publication after registration in clinicalTrials.gov: A cross-sectional analysis. *PLoS Med.*, **6**, e1000144.

Sterling,T.D. (1959) Publication decisions and their possible effects on inferences drawn from tests of significance - or vice versa. *J. Am. Stat. Assoc.*, **54**, 30–34.

Yousefi,M.R. *et al*. (2009) Reporting bias when using real data sets to analyze classification performance. *Bioinformatics*, [Epub ahead of print, doi:10.1093/bioinformatics/btp605, October 21, 2009].