BIOINFORMATICS LETTER TO THE EDITOR Vol. 29 no. 20 2013, pages 2664–2666
doi:10.1093/bioinformatics/btt458

Data and text mining

Advance Access publication August 8, 2013

On representative and illustrative comparisons with real data in bioinformatics: response to the letter to the editor by Smith *et al.*

Anne-Laure Boulesteix

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

Associate Editor: Martin Bishop

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

Received on June 17, 2013; revised on July 23, 2013; accepted on August 2, 2013

1 INTRODUCTION

Smith *et al.* (2013) recently published an interesting letter to the editors of *Bioinformatics* outlining the importance of validation of new proposed algorithms through a thorough comparison with existing approaches. We completely agree with the necessity of more comparison studies in general to help end users to make an informed choice based on objective criteria. We also agree that 'The practical result [of the lack of comparison studies] is that practitioners stop short of exhaustively evaluating all the possible options and choose based on some other criteria (e.g. popularity, ease of use or familiarity)', and that this may harm both algorithm developers and users—as also claimed in one of our previous publications on this topic (Boulesteix *et al.*, 2013a). In a few words, we fully comply with the claim of the authors that everybody (users and algorithm makers) suffers from the trend observed in bioinformatics toward the publication of many new algorithms without proper comparison between these algorithms.

We also agree with Smith *et al.* (2013) that articles suggesting new algorithms should always include a comparison with existing methods. Here we guess that the authors implicitly mean comparisons based on real datasets—as opposed to simulated data that are often investigated in other fields related to data analysis such as statistics. However, our opinion diverges when it comes to defining the goals of such comparison studies and their interpretation. Our major argument is that comparison studies conducted as a part of an article suggesting a new algorithm rarely reach the level of representativity and objectivity required to be used as guidance by other researchers when choosing their algorithm. The two underlying main ideas behind this claim are that (i) comparison studies comparing a new algorithm with existing algorithms are often severely biased for different reasons, as documented in our empirical study based on the example of supervised classification with high-dimensional data (Jelizarow *et al.*, 2010); simply, when reading articles on new methods, most of us think 'well, of course they say their method is better, but…'; and (ii) as stressed by Smith *et al.* (2013), an extensive and fair comparison study requires a lot of work, care and attention in itself (Boulesteix *et al.*, 2013a, b) and can hardly be conducted within a study focused on another problem—the development of a new algorithm. To specify our point of view, we suggest a taxonomy of comparison studies based on real datasets while referring to the ideas presented by Smith *et al.* (2013).

2 REPRESENTATIVE AND ILLUSTRATIVE COMPARISONS

We essentially classify comparison studies based on real datasets into two categories: representative and illustrative comparisons. Representative comparisons aim to give conclusions on the new method for a certain field of application not limited to single datasets. These comparisons include several datasets chosen from a certain field of application that is ideally clearly defined. One hopes from a well-conducted representative comparative application that it will give information guiding the choice of the method in future similar applications from the same field. Note that we do not expect a particular method to outperform all other methods on all datasets from the considered field (Hand, 2006). We rather see representative comparative applications as giving information on the expectation of performance over the 'distribution of distributions' that is characteristic for this field (or, simply, over the datasets from this field); see Boulesteix *et al.* (2013b) for a formal probabilistic framework.

In practice, however, representative comparisons presented in articles introducing new methods have three major pitfalls that often make them fail their goal. The first pitfall is that they are often substantially biased in favor of the new method for different reasons, including optimization of the datasets (Yousefi *et al.*, 2010), of the settings or of the method's characteristics, and publication bias (Boulesteix *et al.*, 2013a; Jelizarow *et al.*, 2010). Another issue that is probably even more difficult to address is the better expertise of the authors on the new method they have been working on for months. As stressed by Smith *et al.* (2013), applying algorithms designed and programmed by other researchers is often far from easy. The focus of a researcher developing new algorithms is on the new algorithms, and he/she can spend only a limited amount of time and energy trying to understand and optimize the use of other algorithms. It is also likely that researchers spend much time solving a problem occurring in their own algorithm, whereas they tend to spontaneously accept inferior results from another algorithm without trying to solve the problem.

In this context, we also claim that it would be extremely difficult (if not impossible) for referees to detect all these issues in

the articles they have to review. A bug in one of the competitive algorithms may perhaps be discovered if the referee spends a few hours or days checking the authors' code. This requires that all codes are made available (which is currently not always the case; see Hothorn and Leisch (2011) for a recent study on this topic) and that the referee has much time to spend on this unpaid reviewing task (this is also unlikely). As far as other aspects such as optimization of the method's characteristics are concerned, there is in our understanding no way even for the most careful referee to identify them with certitude. Based on all these facts, we stressed the importance of neutral comparison studies (i.e. comparisons that are not part of an article suggesting a new algorithm) in a previous publication (Boulesteix *et al.*, 2013a).

The second pitfall is that applications intended as representative comparisons are in practice often underpowered, as documented in Boulesteix *et al.* (2013b) in the special case of supervised learning. By underpowered, we mean that the performance (or the difference between the performances of two algorithms) is so variable across datasets that more datasets would be needed to draw conclusions from a statistical testing perspective. In a literature survey with focus on supervised classification, Boulesteix *et al.* (2013a) found that the median number of datasets considered in comparison studies included in articles on new algorithms was only five. In the statistical framework of Boulesteix *et al.* (2013b), assuming a significance level of 0.05, a power of 0.8 and a 'typical value' of 7% for the standard deviation of the performance difference over the datasets, this number of datasets would only allow to detect as significant a difference in error rates of $>9.5\%$—a big difference! Thus, simply conducting any comparison study is not sufficient to provide guidance: the comparison study also has to have enough power; a condition that is almost never fulfilled in practice.

The third pitfall is that it is difficult to draw datasets at random within a defined field of interest. This problem is characterized by lack of literature and definition problems. It should be addressed in future research. Most importantly, it may be related to the overoptimistic bias mentioned earlier—both because authors might tend to underreport the results obtained with datasets that are not favorable to their new algorithm (Yousefi *et al.*, 2010) and because they might choose datasets that are somehow interrelated and yield a distorted picture of the performance of the new algorithm in the whole field of interest.

Ideally, representative comparisons that aim to yield general conclusions for an application field should roughly follow the rules considered as standard in many substantive fields, for example in clinical research, whereby in our metaphor datasets play the role of patients, methods play the role of treatments and applications fields play the role of populations. This implies, among others, that one (i) considers power issues while designing a comparison study, especially while determining the number of datasets (Boulesteix *et al.*, 2013b); (ii) the selection of the datasets follows systematic and well-documented criteria in the vein of inclusion criteria used in trials; (iii) subgroup analyses are planned previously and interpreted cautiously; (iv) the main outcome (e.g. the error rate in the case of supervised learning) is clearly defined; (v) the datasets are selected at random or at least potential selection biases are discussed; and (vi) dropout (datasets that are disregarded in the course of the study) and its reasons are well-documented.

Note that these recommendations would address the second pitfall but only partially the first one. Following our metaphor, the first pitfall in clinical research would be that inventors of a new treatment tend to be overoptimistic regarding its efficiency for many reasons, a fact that is widely recognized in medical research. In our context, the six recommendations outlined earlier should be completed by a seventh one to better address the first pitfall, say, 'define the method at the beginning of the study and do not adapt it depending of its results on the considered datasets', to avoid overfitting of the new algorithm on the considered datasets (Jelizarow *et al.*, 2010; Rocke *et al.*, 2009).

Considering the three pitfalls discussed above, it is clear that designing a representative comparison study is an extremely difficult and time-intensive task that can and should probably not be performed in all articles presenting new methods. Indeed, many real data applications presented in articles on new methods are meant as examples. This is what we denote as illustrative comparisons. They may demonstrate the use of the new method and point to specific aspects, such as software implementation (possibly including exemplary code as additional file), preliminary data preparation, parameter/variant choice or computation time. They might also show in which form the results are obtained and how to interpret them.

The main concern related to illustrative comparisons is that their results are sometimes (wrongly) interpreted in the literature and discussed as if they were representative. Typically, conclusions are drawn on the superiority of a method (most often the new method; see Boulesteix *et al.* (2013a) for an empirical study on this topic) based on a too small number of datasets that actually do not allow to draw such conclusions from a testing perspective. Coming back to our metaphor, it would be as if a team of medical scientists established the superiority of a new treatment based on only $n=2$ or $n=3$ patients. It just does not make sense. Similarly, it does not make sense to make conclusions such as 'method A performs better than method B for datasets with the characteristics XYZ' if the study is based on, say, one dataset with this characteristic and one without. It would be as if a medical team said that treatment A is more efficient than treatment B for male patients just because it was the case for the considered male patient but not for the considered female patient.

Note that the term 'illustrative' does not necessarily imply that the comparison criteria are qualitative. But it implies that these comparisons and criteria are interpreted as examples, and not as representative of the considered field. In particular, differences in performance should not be interpreted in terms of guidance for the choice of the algorithm—even if the results of illustrative comparisons may provide information on the order of magnitude of the relative performance of the considered algorithms and tell us whether, roughly speaking, the 'new algorithms behave as expected in real data settings'.

The main difference between the two types of comparisons is the way in which datasets are selected that essentially affects their interpretation. In representative comparisons, selection is ideally performed at random within the defined field, and their number is chosen by taking power issues into consideration, whereas in illustrative comparisons datasets are selected because they are

interesting to better present the new method. For example, in an illustrative application it is acceptable to select two extreme datasets, say (in the case of supervised learning), a dataset where the response is easy to predict and a more difficult dataset.

Note that other important aspects of the new algorithms—such as ease of use, speed, conceptual simplicity or generalizability—can be adequately addressed in an article even in the absence of representative comparison study. However, it is important to note that (i) these aspects are of no consequence if the performance turns out to be bad, and (ii) some of them (such as ease of use and speed) also depend on the considered dataset. The above considerations on the selection of example datasets may thus also be relevant to aspects of the algorithms other than performance, even if our classification into illustrative and representative studies primarily focuses on performance.

3 CONCLUSION

In conclusion, we agree that comparisons of new algorithms to existing algorithms are important. As Smith *et al*. (2013), we make a plea for more comparison studies on real datasets in bioinformatics literature. Suggesting an algorithm without even running it on a real dataset is just unacceptable. Going back to our parallel between medical and computational sciences, it would be as if a physician described a new therapy without even showing that it was successful on a few patients. The application of the new algorithm to, say, at least two distinct example datasets with different characteristics should be a minimum non-negotiable requirement for publication. Generally, we also think that in research articles *applying* bioinformatics algorithms to obtain substantive results, the results should not be based on a new novel algorithm that is scarcely described, applied only to the dataset of interest and not compared with any other method.

We also believe that both approaches discussed in this letter—illustrative and representative—may make sense, but that they are completely different and should be reported differently. Each approach implies a different way to report the results and a different way to design the experiment. Reporting and design should be consistent. In practice, the most frequent violation of this principle is when algorithm developers 'feel' that their new method might be better than existing methods based on an underpowered comparison study and report their results as if the comparison was representative. The design of a representative comparison study is so difficult and time consuming and the risk of bias in favor of the new method is in practice so important that most applications presented in articles on new methods should probably be seen as illustrative applications. In our letter, we discussed a few conditions that a good representative application should in our view fulfill to be really representative. However, we believe that much more effort is needed to define precise criteria and guidelines. In particular, the neutrality issue addressed in Boulesteix *et al*. (2013a) should be carefully taken into account.

Referring to the sentence 'these evaluations ought to be primarily provided in the novel algorithm publications themselves' (Smith *et al*., 2013), we again point out that representative comparisons (as defined in our letter) can hardly be performed within an article on a new method, especially because of the well-known optimistic bias in favor of the new method (see also the editorial by Rocke *et al*., 2009). Addressing this general issue related to scientific methodology and publication practice will probably need a long time and much coordinated effort from all parties—researchers, editors and reviewers. This problem is also connected with publication bias and publication of negative research findings (Boulesteix, 2010).

In the meantime, we believe that authors and readers should interpret illustrative comparisons as such—without implicitly assuming that they give information about what would happen on other datasets, and that journals might give more attention to 'neutral' comparison studies entirely devoted to the comparison task itself. This would give a motivation to potential authors of such comparison studies: if they know that their work will not only provide useful information to other scientists but also have good chance to be published in a high-ranking journal, they will be more likely to conduct such a study. Motivating potential authors of comparison studies by publishing more of these studies is certainly easier than motivating overloaded reviewers to spend hours investigating the possible bias of a comparison study included in an article on a new method.

To conclude, in this letter we tried to clarify some aspects left unaddressed by Smith *et al*. (2013). The exact definition of requirements for the publication of new algorithms, however, cannot be formulated within this modest framework. Such guidelines should be the result of coordinated efforts of consortia involving a large number of scientists, similarly to the teams working on reporting guidelines in medical sciences such as the EQUATOR network (Altman *et al*., 2008).

ACKNOWLEDGMENT

I thank Manuel Eugster for helpful comments.

Conflict of Interest: none declared.

REFERENCES

Altman,D.G. *et al*. (2008) EQUATOR: reporting guidelines for health research. *Lancet*, **371**, 1149–1150.

Boulesteix,A.L. (2010) Over-optimism in bioinformatics research. *Bioinformatics*, **26**, 437–439.

Boulesteix,A.L. *et al*. (2013a) A plea for neutral comparison studies in computational sciences. *PLoS One*, **8**, e61562.

Boulesteix,A.L. *et al*. (2013b) A statistical framework for hypothesis testing in real data comparison studies. *Technical Report*, Department of Statistics (LMU), No. 136, http://epub.ub.uni-muenchen.de/14324/ (5 September 2011, date last accessed).

Hand,D.J. (2006) Classifier technology and the illusion of progress. *Stat. Sci*., **21**, 1–14.

Hothorn,T. and Leisch,F. (2011) Case studies in reproducibility. *Brief. Bioinform*., **12**, 288–300.

Jelizarow,M. *et al*. (2010) Over-optimism in bioinformatics: an illustration. *Bioinformatics*, **26**, 1990–1998.

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

Smith,R. *et al*. (2013) Novel algorithms and the benefits of comparative validation. *Bioinformatics*, **29**, 1583–1585.

Yousefi,M.R. *et al*. (2010) Reporting bias when using real data sets to analyze classification performance. *Bioinformatics*, **26**, 68–76.