FREE TO PUBLISH - FREE TO ACCESS - PUBLIC PATRONAGE

Data constraints, bias and the (mis-)use of scientometrics for predicting academic success: A comment on van Dijk et al.

Jan O. Engler[1]✉,[2] & Martin Husemann[3]

[1]Zoological Researchmuseum Alexander Koenig, D-53113 Bonn, Germany
✉ j.engler.zfmk@uni-bonn.de
[2]Department of Wildlife Sciences, University of Göttingen, D-37077 Göttingen, Germany
[3]General Zoology, Institute of Biology, University of Halle-Wittenberg, D-06120 Halle, Germany

Citation: Engler J & Husemann M (2014) Data constraints, bias and the (mis-)use of scientometrics for predicting academic success: A comment on van Dijk et al.. ProcPoS 1:e8

DOI: 10.14726/procpos.2014.e8

What makes a successful researcher? Junior scientists need to start asking this question early if they want to have a perspective in science, leading to the ultimate goal: a tenured position.

In their study van Dijk et al. (2014) tackle this problem and try to identify the factors leading to a principal investigator (PI) position based on publication statistics derived from the PubMed data base. The authors identify the number of publications, the impact factor of the journals they were published in and high numbers of citations as most important contributors for having a successful career in science. While it may help young scientists to know about the different factors leading to success in their field, the study and especially their predicting algorithm may be discouraging specifically for young researchers working in specific fields of science where publications may either be difficult to obtain or high impact factors may not be achievable. The negative influence is enhanced due to problems in the choice of data the authors used, as PubMed does not equally cover all biological disciplines leading to a biased data source that has strong negative effects on the outcome of the predictive model. If young scientists put too much emphasis on such predictors they may be discouraged to pursue a career in science. Further we argue that in general it may be foolish of us to reduce ourselves and our success to a few indices based on a single metric of success, publications.

Van Dijk and colleagues claim that in using their model (a machine learning approach based on a linear model with a five-fold cross-validation scheme and ten iterations, accessible via www.pipredictor.com) every scientist from the biological field can easily calculate their likelihood of becoming a PI. This is not true for a number of reasons: 1) The demands regarding scientific output and certain "quality" metrics such as the Impact Factor (IF) differ strongly between fields (see Stergiou & Lessenich 2013 and references therein) and should therefore not be pooled in a single analysis. 2) Not all scientific fields relevant for biologists are equally well represented in PubMed. Taking these two main methodological issues into account we ask in turn, how representative the authors' data set was for model training. To

exemplify this, we used the ISI Thompson Reuters Journal Citation Report from 2012 and selected different disciplines from biological sciences according to the ISI categories: Biodiversity Conservation, Ecology, Evolutionary Biology, Medicine (General, Internal), Medical Research and Microbiology. While we expect that the latter three categories should be well represented in PubMed as of 1996 (which is the starting year of van Dijk et al.'s data set), the first three were likely not to be. We compare the coverage of the top 10% journals of either category and compared impact factors among the different categories.

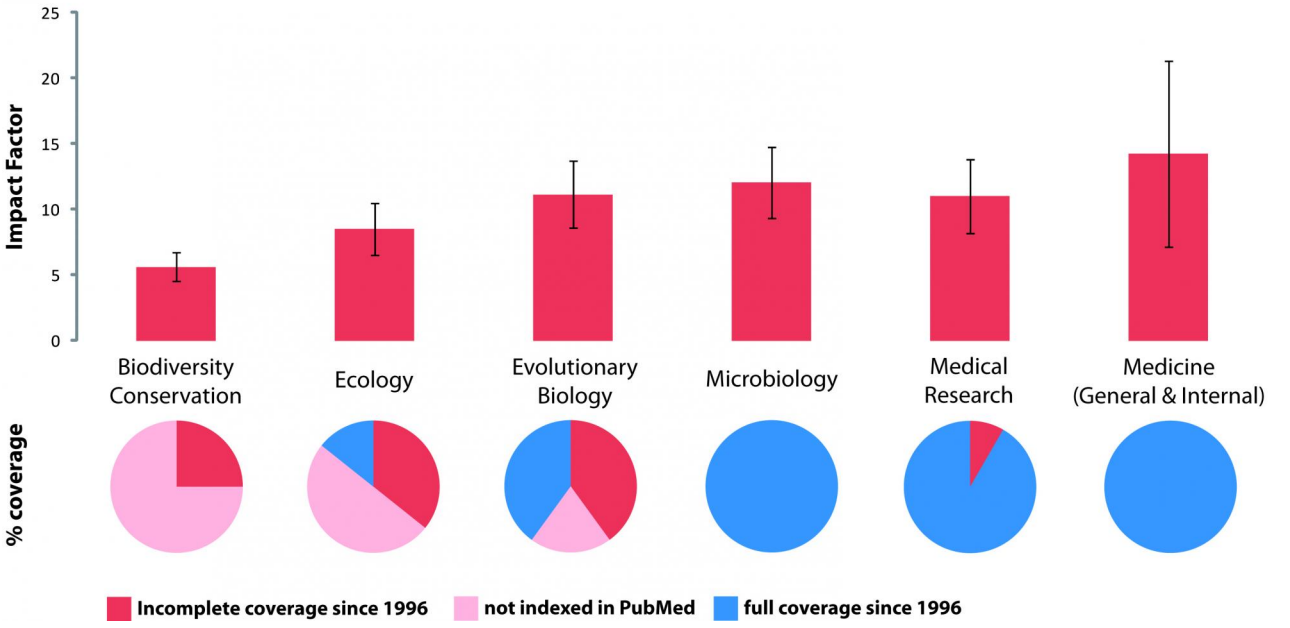

Figure 1. Comparison of the representation of six different sub-disciplines of biology in PubMed. The upper half shows the distribution of the Impact Factors for the year of 2012 (mean and sd) for the top 10% journals for the respective field. The lower half shows the coverage of those journals in the PubMed database since 1996 (the date of starting data compilation by van Dijk et al. 2014). Blue represents a full coverage since 1996, red shows an incomplete coverage since that time. Light red displays the fraction of journals that were not even indexed in PubMed.

We found that medical and microbiological fields were much better represented in PubMed than other biological disciplines (Figure 1). In the less well represented fields, there is not only a high fraction of journals with incomplete coverage (ranging from 25% in biodiversity conservation to 40% in evolutionary biology), but also a large number of journals that are not indexed at all (75% in biodiversity conservation to 20% in evolutionary biology). The lack of indexing in certain biological sub-disciplines is not a big problem per se, when considering the scientific discipline as a co-factor in the model, as it just restricts the applicability of van Dijk et al.'s model to those fields having a full coverage of their respective journals in PubMed. However, the amount of fully and partly covered journals in those sub-disciplines in the PubMed data base leads to an incomplete representation of the authors publication records that are working in these fields; such authors were likely to be included in the compilation of the data base used by van Dijk et al. to train their model. Our own experience with the PIpredictor software revealed that only about 20% of our ISI ranked publications were found in PubMed and/or used by the PIpredictor software. However, given the selection criteria of van Dijk et al.'s database compilation and PI determination ('We included only authors whose first publication was between 1996 and 2000. […] we consider as becoming PI only those authors that have at least three last author publications and measure the time to PI as the time between that person's fist publication and the time of the second last-author publication […].', van Dijk et al. 2014 - supplemental information), such incomplete PubMed author profiles would have been fully considered in the database and by the model, if matching the relevant

time frame (i.e. first publication between 1996 and 2000). The inclusion of these author profiles leads to a severe bias in the dataset and consequently to wrong and likely misleading conclusions derived by the predictive model that affects the general meaning of the outcome, even for the well covered medical and microbial sub disciplines.

Authors with incomplete publication records are more likely to be selected into the group of "unsuccessful" scientists (i.e. not reaching two last authored publications).. Depending on the amount of unconsidered publications due to a lack of coverage of the respective sub discipline, scientists from these fields have to publish more last author publications to meet the criteria and be considered a PI (e.g. twice as many if publication coverage of the respective author by PubMed is just 50%). In addition, this bias likely also affects other metrics such as the time a scientist needs to become a PI (i.e. time between the author's first publication and its second last-authored publication).

This bias might also explain the divergence in some of the starting trajectories between 'PI' and 'non PI' scientists (cf. van Dijk et al. Fig. 1 D-G). It is somewhat surprising that researchers that later become PI already start with a higher publication rate or average IF in their first year of their career. Under random starting conditions we would expect that later PI's and non-PI's would start publishing in the same journals with a similar rate as everyone faces similar problems in the very beginning of their academic career. The reasons for this pattern were not discussed in the article, but the finding could likely be the result of the incomplete sampling and different IF in different sub-disciplines as described above; alternatively the result may even represent a true signal, i.e. PI's originate from better performing labs than non PI's. This would be worth investigating in the future.

Impact factors themselves have been a topic of many recent discussions. The criticism of IF and its relatives has been based on the sky-rocketing IF of new online journals, the great divide between journal rankings in different sub-disciplines (see Fig. 1) and the misuse as a measure of individual performance (e.g. Stergiou & Lessenich 2013). While this is a problem for scientists in some of the lower ranked disciplines, the bigger problem might be the emphasis scientists put on these numbers. Here we would like to cite Stergiou & Lessenich who write that 'higher education must urgently take control of its own metrics'. The study by van Dijk et al. raises the question of solely focusing on metrics in an unintended way. While Impact Factors may be a simple metric to rank journals and scientists it may negatively affect science itself. Several papers have addressed this topic already e.g., Brembs et al. 2013, Marks et al. 2013, or meaningful blogs such as 'sick of impact factors' (retrieved from: http://occamstypewriter.org/scurry/2012/08/13/sick-of-impact-factors/) or 'the impact of impact factors' by Graham Davey (retrieved from: http://www.papersfromsidcup.com/graham-daveys-blog/the-impact-of-impact-factors-good-for-business-but-bad-for-science). Further, the 'tyranny' of top tier journals has already led to a boycott by recent Nobel Price winner Randy Schekman (retrieved from: http://www.theguardian.com/science/2013/dec/09/nobel-winner-boycott-science-journals). This should make us think! On the other hand, a person who already has won a Nobel prize is no longer under the pressure young researchers are, to ultimately become PI and find a permanent position in science. Therefore, young scholar cannot afford such steps as the system pushes them to maximize their metrics. In course of this 'race' the quality of science may suffer (Marks et al.2013).

In general, a researcher that does good science will reach better journals automatically and more readers will cite the respective studies. However, a bad scientist may just focus on such metrics and optimize it disregarding the ethics of science and the scientific progress. Hence, instead of promoting the spread of good scientific work, a too strong focus on publication metrics may lead to the opposite in the course of a 'struggle for survival' situation. Further, more substantial long-term studies may be inhibited by encouraging researchers to publish as frequently as possible. This, again, would lead to a decrease in scientific quality. Hence, it may be preferential for the scientific community to refocus on the essentials and, instead of focusing on achieving high metrics, doing sound and solid research. Especially young scientists should be granted this opportunity rather than being constantly pushed towards higher publication rates (Laurance et al. 2013, Laurance et al. 2014, Bruna 2014), currently the only way to receive funding and obtain a tenured position. Given the high popularity of the PIpredictor software (e.g. featured several times in the leading academic journals Nature and Science, and numerous other

news media, >57.000 PI predictions as of 13th Oct. 2014) it is crucial to mention the flaws and shortcomings of this approach, in both a technical and an ideological manner.

References

Brembs B, Button K, Munafo M (2013) Deep impact: unintended consequences of journal rank. Frontiers in Human Neuroscience 7: 291. http://dx.doi.org/10.3389/fnhum.2013.00291

Bruna EM (2014): On identifying rising stars in ecology. BioScience 64: 169. http://dx.doi.org/10.1093/biosci/biu003

Laurance WF, Useche DC, Laurance SG, Bradshaw JA (2013): Predicting publication success for Biologists. BioScience 63: 817-823. http://dx.doi.org/10.1525/bio.2013.63.10.9

Laurance WF, Useche DC, Laurance SG and Bradshaw JA (2014): Identifying rising stars in biology: A response to Bruna. BioScience 64: 169-170. http://dx.doi.org/10.1093/biosci/biu006

Marks MS, Marsh M, Schroer TA, Stevens TH (2013) Misuse of journal impact factors in scientific assessment. Traffic 14: 611-612. http://dx.doi.org/10.1111/tra.12075

Stergiou KI, Lessenich S (2013): On impact factors and university rankings: from birth to boycott. Ethics in Science and Environmental Politics 13 http://dx.doi.org/10.3354/esep00141

van Dijk D, Manor O, Carey LB (2014): Publication metrics and success on the academic job market. Current Biology 24: R516-R517. http://dx.doi.org/10.1016/j.cub.2014.04.039

Funding: No specific funding is attributed for this work

Competing interests: Authors declare no competing interests exist