Publication metrics and success on the academic job market

David van Dijk[1,4], Ohad Manor[2,4], and Lucas B. Carey[3,*]

The number of applicants vastly outnumbers the available academic faculty positions. What makes a successful academic job market candidate is the subject of much current discussion [1–4]. Yet, so far there has been no quantitative analysis of who becomes a principal investigator (PI). We here use a machine-learning approach to predict who becomes a PI, based on data from over 25,000 scientists in PubMed. We show that success in academia is predictable. It depends on the number of publications, the impact factor (IF) of the journals in which those papers are published, and the number of papers that receive more citations than average for the journal in which they were published (citations/IF). However, both the scientist's gender and the rank of their university are also of importance, suggesting that non-publication features play a statistically significant role in the academic hiring process. Our model (www.pipredictor.com) allows anyone to calculate their likelihood of becoming a PI.

In order to quantify precisely if and when individual authors will become a principal investigator (PI) we generated a set of 25,604 uniquely identifiable authors (Figure 1A and Supplemental information). We then quantified more than 200 different metrics of publication output for authors who became PIs and for those who didn't. We find that whether or not a scientist becomes a PI is largely predictable by their publication record (Figure 1B,C), even taking into account only the first few years of publication (Figure 1D–G). In order to quantify the effect of each publication feature independent of other confounding variables, we developed a statistical model (Supplemental information). This model is able to predict with relatively high accuracy who becomes a PI (held-out test AUC = 0.83), and how long this will take (R^2 = 0.38). We note that a minimal model that uses only the five most predictive features still has significant predictive power (AUC = 0.74; Supplemental information).

As expected, authors with more first author publications, and with more papers in high impact factor journals, are more likely to become PIs (Figure 1B). In addition, they have a higher *h*-index (*h* papers with at least *h* citations each), consistent with the idea that current *h*-index is predictive of future scientific success[4]. However, the actual number of citations is less predictive of becoming a PI than journal impact factor (Figure 1B), suggesting that currently, the perceived quality of a publication (*i.e.*, journal impact factor) is given more weight than its actual quality (*i.e.*, number of citations). Because the number of citations a publication will receive is correlated with the impact factor of the journal, we examined the number of citations divided by the impact factor (cites/IF). We find that in a linear model, cites/IF is the fourth most predictive feature after impact factor, number of publications and gender (Figure 1B; Supplemental information). This suggests that hiring committees also take into account exceptional papers published in lower impact factor journals.

We found that many scientists who will become PIs never published in high impact factor journals. In order to better understand how these authors manage to become PIs we analyzed separately the group of authors who become PIs but have very low impact factor publications (lower than 75% of all non-PI authors). We find that these authors have a two-fold increase in their first-author publication rate compared to authors who do not become PIs, suggesting that more first-author publications per year can compensate for lack of high impact factor publications.

While authors with more first or second author publications are more likely to become PIs, we find that more middle (non-first and non-second) author publications are of no help unless they are published in high impact journals (Figure 1B). In addition, authors who are middle author on papers with many co-authors are less likely to become PIs (Figure 1B). While staff scientists and technicians may cause much of this effect, it still holds for the first author. The small negative correlation between the number of co-authors and the probability of becoming a PI suggests that first authors on papers with many co-authors are given less credit for these publications.

By almost all metrics, PIs differentiate themselves from authors that eventually will leave academia in the first years of their career (Figure 1D–G). However, while around half of authors become PIs less than seven years following their first publication, short pre-PI career scientists show different publication behavior than longer pre-PI career authors. Authors who take longer than seven years to become a PI have more citations per paper than authors who become PIs more quickly (Figure 1F), suggesting that scientists who publish important papers in low impact factor journals can still become PIs, but that this route takes more time.

The set of PIs is highly enriched for scientists who attended higher ranked universities, and university ranking is highly correlated with many other features. However, we find that university rank is predictive of becoming PI independent of other publication features (university rank adds, on average, 0.04 to the AUC in cross-validation, t-test $p < 0.01$). In addition we find that PIs, but not non-PIs, increase their university ranking (as ranked in the Shanghai Jiao Tong Top 500 research universities) in the first five years of their careers (Figure 1G), suggesting that they do their postdoc (or collaborative work) at a university that is better than the one in which they completed their PhD. A decline in the mean university rank among future PIs with longer careers suggests that, on average, scientists from higher ranked institutions become PIs before scientists from lower ranked institutions (Figure 1G).

Men are overrepresented as PIs, yet even after correcting for all other publication and non-publication derived features, being male is positively predictive of becoming a PI (increase in AUC = 0.02, t-test $p < 0.01$). Given the same publication record, men are more likely than women to become PIs.

This is the first study that quantifies what is predictive of an academic career in terms of becoming a principal investigator. While the journal impact factor and the number of publications are the most predictive features, the data suggest that outstanding work will be noticed, regardless of the impact factor of

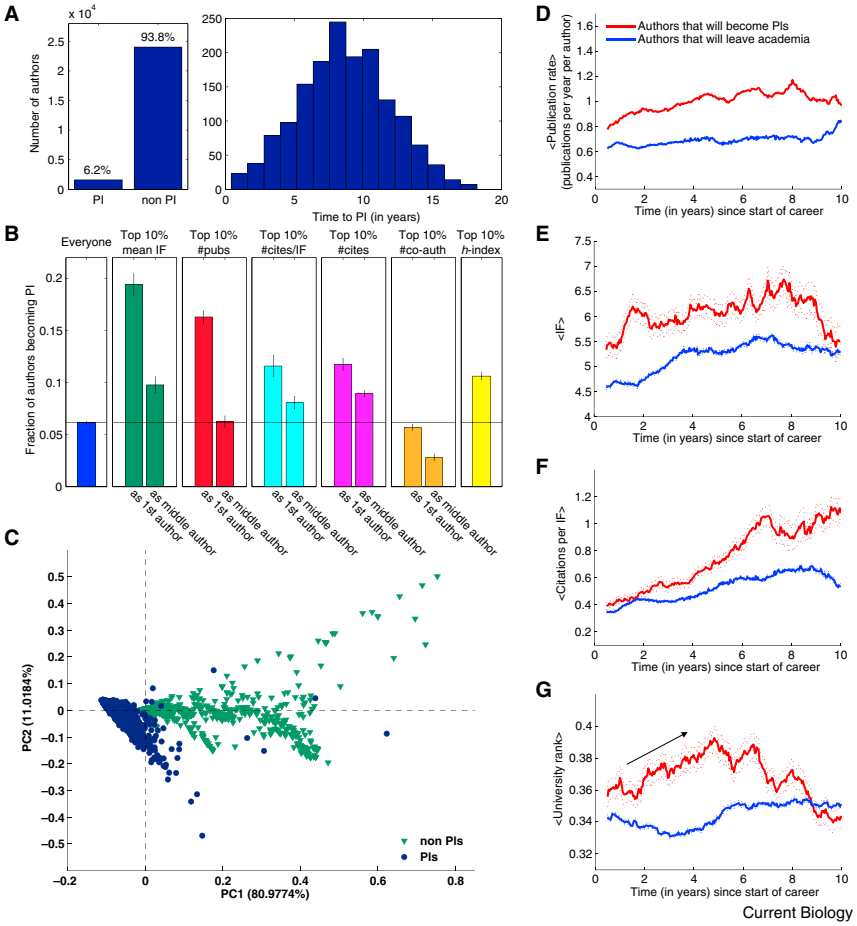

Figure 1. Publication features, prior to becoming a PI or leaving academia, accurately separate future PIs from non-PIs.
(A) In our data set of 25,604 authors, 1583 (6.2%) become a PI (A, left). (A, right) Histogram of the distribution of time to PI for all authors that become PIs. (B) Shown are publication features that separate PIs from non-PIs. The blue bar shows the total fraction (6.2%) of authors that become PIs. Green, red, cyan, magenta, orange and yellow bars show the fraction of authors in the top 10% for a given grouping that becomes a PI. Error bars show the standard deviation for this calculated fraction following 100 bootstraps. For each author, only non-last author publications prior to becoming a PI are used for the calculation. (B, cyan bar) Authors are grouped according to the mean number of citations per IF (journal impact factor) for publications in which they are first author (left cyan bar) or middle author (right cyan bar). (B, magenta bar) Authors are grouped according to the mean number of citations for publications in which they are first author (left magenta bar) or middle author (right magenta bar). (B, orange bar) Authors are grouped according to their average number of co-authors, for papers in which they are either first (left yellow bar) or middle (right yellow bar) author. (B, yellow bar) Authors are grouped according to their *h*-index. (C) Principal component analysis is shown in which the first two principal components explain 92% of the variance. Future PIs are shown in blue circles, future non-PIs in green triangles. (D–G) Shown are the trajectories of various publication features in time, for authors that will eventually become PI and for authors that will eventually leave academia. Dotted lines are error-bars obtained by bootstrapping. Authors who will eventually become PI (red lines) show, on average (compared to authors who will eventually leave academia, blue lines), already in early career, an increased rate of publication (D, mean publication rate in time), and an increased journal impact factor (E, mean IF in time). Authors that have longer pre-PI careers show an increased number of citations per IF (F, mean number of citations per IF in time). Authors who will eventually become PI go to higher ranked universities (G, mean university rank in time). In addition, for authors that will become PI, university rank appears to increase within the first 5 years of their careers (G, arrow).

the journal in which it is published. Perhaps surprisingly, the number of co-authors has a slight negative effect. Indeed, measures of scientific impact that take co-authorship into account may be preferred [5,6]. The *h*-index has significant predictive power of becoming a PI, supporting previous findings in which *h*-index was able to predict future scientific success [4]. Better universities attract better people and it is therefore expected that they produce more PIs. But we found that this effect persists even after correcting for publication success. Either university rank correlates with some non-publication features (e.g. 'soft' skills) or names of highly ranked universities look good on applicants' CVs. In addition, we find a bias in favor of men who come from highly ranked universities, but cannot differentiate bias in the hiring process from a self-selective one in which men from high ranked universities prefer to become PIs. In addition, our model measures correlation, not causation. Our results suggest that currently, journal impact factor and academic pedigree are rewarded over the quality of publications, which may dis-incentivize rapid communication of findings, collaboration and interdisciplinary science.

Supplemental Information

Supplemental Information including experimental procedures, one figure and one table can be found with this article online at http://dx.doi.org/10.1016/j.cub.2014.04.039.

Acknowledgements

We thank Eran Segal for comments on the manuscript and suggestions with analysis, and for support during the writing of this manuscript.

References

1. Schmid, S.L. (2013). Beyond CVs and Impact Factors: An Employer's Manifesto | http://sciencecareers.sciencemag.org/career_magazine/previous_issues/articles/2013_09_03/caredit.a1300186
2. Alberts, B. (2013). Impact factor distortions. Science *340*, 787–787.
3. Kirschner, M. (2013). A perverted view of "impact". Science *340*, 1265.
4. Acuna, D.E., Allesina, S., and Kording, K.P. (2012). Future impact: Predicting scientific success. Nature *489*, 201–202.
5. Hirsch, J.E. (2005). An index to quantify an individual's scientific research output. Proc. Natl. Acad. Sci. USA *102*, 16569–16572.
6. Stallings, J., Vance, E., Yang, J., Vannier, M.W., Liang, J., Pang, L., Dai, L., Ye, I., and Wang, G. (2013). Determining scientific impact using a collaboration index. Proc. Natl. Acad. Sci. USA *110*, 9680–9685.

[1]Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot 76100, Israel. [2]Department of Genome Sciences, University of Washington, Seattle, WA, 98195, USA. [3]Department of Experimental and Health Sciences, Universitat Pompeu Fabra (UPF), E-08003 Barcelona, Spain.
[4]These authors contributed equally to this work.
*E-mail: lucas.carey@upf.edu