



Can we trust peer-reviewed science?

Pouvons-nous nous fier à la science évaluée par les pairs?

Carlton Gyles

Discussion is heating up regarding honesty and effectiveness in our system of doing science and communicating it. Refereed journal publications are the main mechanism by which research output is evaluated and presented or not presented to the world and individuals or groups frequently assess how well the system is working. These assessments assume greater importance when the research environment and the methods of journal publication have undergone major changes. Two relevant changes are the massive increase in the number of researchers and the proliferation of for-profit, online only, open access (OA) publishers. As an example, 1 publisher in this category started in 2010 but lists 120 journals on its website. An academic librarian, Jeffrey Beall (1) has a website that lists “questionable, scholarly open-access publishers,” which are considered to be probably predatory; there are presently 471 journals on this list.

Peer review is central to the assessment of quality of research that is published in scientific journals. However, a recent article in *The Economist* (2) cites evidence that the system is not working well. The authors of this article believe that there are problems of poorly conducted studies, unsatisfactory analysis, and inadequate review. One piece of evidence cited refers to reports of failure of independent researchers to replicate studies published in refereed journals. They noted that researchers at an American pharmaceutical company were able to reproduce the published results in only 6 of 53 important basic science studies on cancer and that other researchers at a German pharmaceutical company could reproduce results in only about 25% of 67 major studies related to drug discovery. Concerns were also raised about weaknesses in statistical analyses of data, a common problem being inadequate statistical power. For example, a power of 0.8 is accepted to be a reasonable basis for testing hypotheses, but in one field the typical statistical power achieved by researchers was said to be 0.21, and in another, 0.35. Low statistical power increases the chances of a conclusion that there is no difference between the groups being examined based on

Le débat s’anime à propos de l’honnêteté et de l’efficacité de notre système de réalisation des travaux scientifiques et des communications qui en découlent. Les publications dans les revues évaluées par un comité de lecture représentent le principal mécanisme par lequel les résultats de la recherche sont évalués et présentés ou non présentés au monde et les personnes ou les groupes évaluent fréquemment le fonctionnement du système. Ces évaluations revêtent une plus grande importance lorsque le milieu de recherche et les méthodes de publication des revues ont subi des changements majeurs. Deux changements pertinents sont la hausse massive du nombre de chercheurs et la prolifération d’éditeurs à but lucratif, uniquement en ligne et à libre accès. À titre d’exemple, un éditeur dans cette catégorie qui a lancé ses activités en 2010 présente maintenant 120 revues sur son site Web. Un bibliothécaire universitaire, Jeffrey Beall (1), offre un site Web qui indique «les éditeurs douteux à libre accès universitaire» qui sont probablement considérés comme des éditeurs prédateurs; il figure actuellement 471 revues sur cette liste.

L’évaluation par les pairs se trouve au cœur de l’évaluation de la qualité de la recherche qui est publiée dans les revues scientifiques. Cependant, un article récent publié dans *The Economist* (2) cite des preuves indiquant que le système éprouve des ratées. Les auteurs de cet article croient qu’il existe des problèmes liés à des études mal réalisées, à des analyses insatisfaisantes et à des évaluations inadéquates. Un élément de preuve cité signale des rapports d’échecs de la répétition d’études publiées dans les revues avec comité de lecture. Les auteurs font remarquer que les chercheurs d’une compagnie pharmaceutique américaine ont pu reproduire les résultats publiés dans seulement 6 de 53 études de sciences fondamentales importantes sur le cancer et que d’autres chercheurs d’une compagnie pharmaceutique allemande pouvaient reproduire les résultats d’environ seulement 25 % de 67 études majeures reliées à la découverte de médicaments. Des préoccupations

Use of this article is limited to a single copy for personal study. Anyone interested in obtaining reprints should contact the CVMA office (hbroughton@cvma-acmv.org) for additional copies or permission to use this material elsewhere.

L’usage du présent article se limite à un seul exemplaire pour étude personnelle. Les personnes intéressées à se procurer des réimpressions devraient communiquer avec le bureau de l’ACMV (hbroughton@cvma-acmv.org) pour obtenir des exemplaires additionnels ou la permission d’utiliser cet article ailleurs.

the statistical results ($P > 0.05$), when in fact a real difference does exist and there was inadequate statistical power to detect it.

Inadequate peer review is blamed for much of the perceived state of science publication. *The Economist* (2) placed considerable emphasis on a sting operation carried out by science reporter John Bohannon. He created a fake paper with serious flaws, including lack of controls, and submitted it to 304 peer reviewed, open access journals, nearly half of which were on Beall's list of disreputable science journals. The fact that 157 of the journals accepted the article was used as a basis for decrying the state of peer review in open access science journals. The Directory of Open Access Journals (DOAJ) (3) drew 3 conclusions from the sting operation: "there are predatory publishers, there are problems with peer review, some of the exposed journals are indexed in DOAJ." In response, the DOAJ removed 114 journals from its directory.

Several of the points raised in *The Economist* are worth closer examination. First, let us consider the matter of irreproducibility of experiments. These failures must be due to either fraud or inability to precisely reproduce the studies (if a study is truly replicated then the results must be the same as in the original). It is difficult to believe that 75 to 89% of the studies investigated by the 2 drug companies were fraudulent. It is more likely that inability to obtain the published results of the study is most frequently due to difficulties in replication of experiments (4). It is amazing how some unrecognized aspect of the replication attempt that is not identical to that in the original can lead to a different result.

Next is the matter of poor statistical power. *The Economist* certainly has a point here — many published studies have low statistical power. Researchers cite limitations imposed by funding and by unavailability of cases as reasons for low statistical power. Nonetheless, this is an area that needs to be addressed by researchers, although it is worth noting that many studies do not require statistical analysis.

Then there is the matter of the weak peer review system. The overall poor results of peer review (apparently no peer review in some cases) are highlighted in the Bohannon sting operation. This is partly due to the journals that were selected but it has also been shown that non-open-access journals with high reputation and very high impact factor, such as *Science*, also occasionally publish and highlight terribly flawed studies. Reasons suggested for this state of affairs include: journals favoring what appears to be dramatic new findings, lack of interest in funding or publishing studies that replicate previous work, overworked academic reviewers, increasing specialization resulting in small pools of experts, reliance on the assessment of only 2 or 3 persons, the rise of journals that will publish almost anything for a fee. Interestingly, shoddy work often does get exposed after publication.

Sensationalism has resulted in exaggeration of the weaknesses of peer review, which appears to work well in most cases. However, there are ways in which our present peer review system can be improved. Some journals have put in place protocols designed to do just that. Others are replacing pre-publication review with post-publication review. I will discuss these developments in my next editorial.

ont aussi été soulevées à propos des faiblesses de l'analyse statistique des données et l'efficacité statistique inadéquate est un problème courant. Par exemple, une efficacité de 0,8 est acceptée comme étant un fondement raisonnable pour l'évaluation des hypothèses, mais, dans un domaine, l'efficacité statistique réalisée par les chercheurs aurait été de 0,21, et, dans un autre, de 0,35. Une faible efficacité statistique accroît les risques de tirer la conclusion qu'il n'y a aucune différence entre les groupes examinés en se fondant sur les résultats statistiques ($P > 0,05$) lorsqu'en fait, il existe une différence réelle et que l'efficacité statistique n'a pas pu la déceler parce qu'elle était inadéquate.

On jette le blâme sur un examen inadéquat par les pairs pour une grande part de l'état perçu des publications scientifiques. *The Economist* (2) a fait grand cas d'une opération surveillée exécutée par le reporter scientifique John Bohannon. Il a créé un faux article comportant des erreurs graves, dont l'absence de contrôles, et l'a soumis à 304 revues à comité de lecture et à libre accès, dont près de la moitié se trouvaient sur la liste de Beall de revues scientifiques louches. Le fait que 157 des revues aient accepté l'article a été utilisé pour dénoncer l'état de l'évaluation par les pairs dans les revues scientifiques à libre accès. Le Directory of Open Access Journals (DOAJ) (3) a tiré trois conclusions de cette opération de surveillance : «il y a des éditeurs prédateurs, il existe des problèmes avec l'évaluation par les pairs et certaines des revues mises au jour sont répertoriées dans le DOAJ». Pour répondre à cette situation, le DOAJ a supprimé 114 revues de son répertoire.

Plusieurs points soulevés dans *The Economist* méritent un examen approfondi. Examinons d'abord la question de la reproductibilité des expériences. Ces échecs doivent être attribuables soit à la fraude ou à l'incapacité de reproduire précisément les études (si une étude est vraiment reproduite, alors les résultats doivent être les mêmes que ceux de l'étude originale). Il est difficile de croire que de 75 à 89 % des études examinées par les deux compagnies pharmaceutiques étaient frauduleuses. Il est plus probable que l'incapacité d'obtenir les résultats publiés de l'étude soit le plus souvent attribuable aux difficultés inhérentes à la répétition des expériences (4). Il est incroyable de voir comment un aspect non reconnu particulier d'une tentative de répétition qui n'est pas identique à celui de l'étude originale peut produire un résultat différent.

Vient ensuite la question de la faible efficacité statistique. *The Economist* présente certes un argument valable ici — beaucoup d'études publiées ont une faible efficacité statistique. Les chercheurs citent les limitations imposées par les fonds et l'indisponibilité de cas comme des raisons pour une faible efficacité statistique. Néanmoins, c'est un domaine qui doit être abordé par les chercheurs, même s'il vaut la peine de signaler que beaucoup d'études n'exigent pas d'analyses statistiques.

Puis, il y a la question de la faiblesse du système d'évaluation par les pairs. Les mauvais résultats généralisés de l'évaluation par les pairs (et, dans certains cas, de l'absence d'une évaluation par les pairs) sont mis en lumière dans l'opération de surveillance de Bohannon. Cette situation est en partie attribuable aux revues qui ont été choisies, mais il a aussi été démontré que les revues sans libre accès avec une bonne réputation et un facteur

References

1. Beall J. List of publishers: Beall's List. Available from: <http://scholarlyoa.com/publishers/> Last accessed December 14, 2013.
2. Anonymous. Unreliable research: Trouble at the lab. *The Economist*. Available from: <http://www.economist.com/news/briefing/21588057-scientists-think-science-self-correcting-alarming-degree-it-not-trouble> Last accessed December 14, 2013.
3. Second response to the Bohannon article 2013-10-18. Available from: <http://www.doaj.org/doi?func=news&nId=317&uiLanguage=en> Last accessed December 14, 2013.
4. Meagher P. Why the Economist is wrong about science reproducibility. Available from: <http://www.jove.com/blog/2013/10/21/the-medium-is-the-massage-what-the-economist-failed-to-mention-on-reproducibility> Last accessed December 14, 2013. ■

d'impact très élevé, dont Science, publient et présentent parfois aussi des études terriblement imparfaites. Les raisons suggérées pour cette situation incluent notamment : les revues favorisant ce qui semble être de nouveaux résultats spectaculaires, l'absence d'intérêt pour le financement ou la publication d'études qui répètent des travaux antérieurs, des évaluateurs universitaires surmenés, une spécialisation grandissante causant le regroupement de petits groupes d'experts, l'appui sur l'évaluation de seulement deux ou trois personnes et la montée des revues qui publient presque n'importe quoi moyennant des frais. Fait intéressant, le travail de mauvaise qualité est souvent mis au jour après la publication.

Certes, le sensationnalisme s'est traduit par une exagération des faiblesses de l'évaluation par les pairs, qui semble bien fonctionner dans la plupart des cas. Cependant, il y a des façons d'améliorer le système d'évaluation par les pairs utilisé actuellement. Certaines revues ont mis en place des protocoles

conçus à cet effet. D'autres procèdent au remplacement de l'examen avant la publication par un examen après la publication. Je discuterai ces éléments nouveaux dans mon prochain éditorial.

Renvois

1. BEALL, J. *List of publishers: Beall's List*. Disponible au <http://scholarlyoa.com/publishers/> Dernière consultation le 14 décembre 2013.
2. ANONYME. «Unreliable research: Trouble at the lab», *The Economist*. Disponible au : <http://www.economist.com/news/briefing/21588057-scientists-think-science-self-correcting-alarming-degree-it-not-trouble> Dernière consultation le 14 décembre 2013.
3. *Second response to the Bohannon article 2013-10-18*. Disponible au <http://www.doaj.org/doi?func=news&nId=317&uiLanguage=en> Dernière consultation le 14 décembre 2013.
4. MEAGHER, P. *Why the Economist is wrong about science reproducibility*. Disponible au <http://www.jove.com/blog/2013/10/21/the-medium-is-the-massage-what-the-economist-failed-to-mention-on-reproducibility> Dernière consultation le 14 décembre 2013. ■