The Leiden Manifesto for research metrics

研究計量に関するライデン声明

The Leiden Manifesto for research metrics, Nature, 2015, 520(7548), 429-431 (23 April 2015) by Diana Hicks, Paul Wouters, Ludo Waltman, Sarah de Rijcke & Ismael Rafols. http://dx.doi.org/10.1038/520429a

Translated by Natsuo ONODERA & Masatsura IGAMI (National Institute of Science and Technology Policy, Japan)

科学の管理へのデータの利用が進んでいる。かつてカスタムメードで同分野の研究者(ピア)によって実施されていた研究評価は、今日では日常化し計量に依存している [1]。問題は、現状では評価が、判定よりもむしろデータによって主導されていることである。計量は急激に増加している。通常、善意で行われているが、必ずしもよく理解されておらず、しばしば誤って適用される。グッドプラクティスと解釈についての知識がなく、それについての助言も得られない機関によって評価がますます実施されるのに伴い、システムを改善するために設計されたツールそのものでシステムを損なうという危険が生じている。

2000 年以前には、Institute for Scientific Information (ISI)から提供される Science Citation Index の CD-ROM があり、専門的分析のためにその道のプロによって用いられていた。[ISI を吸収した]Thomson Reuters は、2002 年に統合的 web プラットフォームを開設し、それによって Web of Science データベースが広く利用できるようになった。これと競合する引用索引データベースとして Elsevier の Scopus（2004 年リリース）と Google Scholar（2004 年にベータ版をリリース）が作られた。研究機関の生産性とインパクトを容易に比較できる web ベースのツールである InCites（Web of Science を利用）、SciVal（Scopus を利用）や、Google Scholar を用いて個々の引用プロファイルを分析するためのソフトウェア（Publish or Perish、2007 年リリース）が登場した。

2005 年にカリフォルニア大学サン・ディエゴの物理学者 Jorge Hirsch は *h* 指数を提案し、これによって個々の研究者を対象とした引用の計量化が普及した。雑誌のインパクトファクターへの関心は 1995 年以降着実に高まった。

最近、[科学論文の]社会での利用とオンラインでのコメントに関する計量が勢いを得ている。即ち、F1000Prime が 2002 年に、Mendeley が 2008 年に、Altmetric.com（Nature Publishing Group を所有する Macmillan Science and Education が支援）が 2011 年にスタートした。

我々は、科学計量の専門家、社会科学者、リサーチ・アドミニストレーターとして、科学的パフォーマンスの評価への指標の誤った適用の蔓延に対して、懸念を増大させながら警告を鳴らし続けてきた。以下に述べるのは、多くの事例のほんの一部である。世界中で大学は世界的な大学ランキング（上海ランキングや Times Higher Education リストなど）で自分がどの位置にあるかに汲々としている、それらのリストは（我々の見方では）不正確なデータ(inaccurate data)と恣意的な指標(arbitrary indicators)に基づいているにも拘わらず。

応募者に *h* 指数の値を提出させる採用者もある。*h* 指数の値や「高インパクト」誌への掲載論文数に基づいて昇進を決定する大学もある。特に生医学の分野では、研究者の履歴書(CV)がこれらのスコアを喧伝する場となりつつある。いたるところで、指導教員は博士課程の学生に、彼らが十分

な準備をできていないうちに、高インパクト誌に発表し、外部資金を獲得することを求める。

スカンジナビア諸国や中国では、数字に基づいて研究資金や報奨金を配分する大学がある。たとえば、高業績研究費(performance resources)の配分のため個人のインパクトスコアを計算したり、インパクトファクターが 15 より高い雑誌に発表した研究者に報奨金を与えたりするなどである[2)]。

多くの場合、研究者と評価者は今でも均衡のとれた判定をしている。しかし、研究計量の誤用は無視できないほど拡がっている。

そのため、我々はここにライデン声明を公表する。この名称は、この文書が結実した会議に因んだものである(http://sti2014.cwts.nl を参照)。ここに挙げている 10 の原則のそれぞれは、科学計量の専門家にとって新しいものではないが、これまで条件が整っていなかったため、誰もその全体をまとめることはできていなかった。Eugene Garfield(ISI の創設者)などこの分野の先達は、これらの原則のいくつかについて言及している[3,4)]。しかし、関連の方法に関する専門知識がない大学管理者に評価者が結果を報告する際、それらについて述べる機会はなかった。評価について論ずる文献を探す研究者は、それらの資料が、彼らにとっては接することが少ない無名の雑誌に散在していることに気づかされる。

我々はここに、研究者は評価者に責任を取らせることができ、評価者は彼らの指標に責任を取らせることができるように、計量に基づく研究評定におけるベストプラクティスの蒸留物(the distillation)を提示する。

ライデン声明－10 の原則

原則 1 定量的評価は、専門家による定性的評定の支援に用いるべきである。

定量的計量は、ピアレビューで生じやすいバイアスについて異なる見方を提示し、考察を深めるのに役立つ。同業研究者について判定することは広範な関連情報なしには難しいので、これによりピアレビューは強化されるはずである。しかしながら、評定者は意思決定を数字に任せてはならない。指標は情報に基づく判定を代替してはならない。評定者はそれぞれが行う評定に責任を保持している。

原則 2 機関、グループ又は研究者の研究目的に照らして業績を測定せよ。

プログラムの目標はその開始時に明示されるべきであり、また、業績を評価する指標は、それらの目標と明確に関係づけるべきである。指標の選択やその活用に際しては、より幅広い社会経済的及び文化的な状況を考慮すべきである。科学者の研究目的は様々である。学術的知識の最前線を進める研究と、社会的問題の解決を目指す研究とは目的が異なる。学術的なアイディアの卓越性よりも、政策、産業、あるいは公衆への貢献に基づく評価もある。すべての状況に適用できる単一の評価モデルはない。

原則 3　優れた地域的研究を保護せよ。

世界の多くの地域で、優れた研究は英語で発表されると見なされている。たとえば、スペインの法律は、同国の学者が高インパクトの雑誌に発表することを望ましいとしている。インパクトファクターは、米国中心で、いまだにほとんどが英語である Web of Science 収録の雑誌を対象に計算されている。こうしたバイアスは、国・地域についての研究が多い人文・社会科学において特に問題が大きい。他の多く分野でも、国・地域という側面を持つ。例えば、サハラ以南アフリカにおける HIV の疫学などの例がある。

しかし、このような多元性や社会的関連性は、高インパクトのゲートキーパーたる英語雑誌の関心を得るような論文を創出するために抑制される傾向がある。Web of Science で高引用を得ているスペインの社会学者たちは、抽象モデルに長年取組んでいるか、米国データの研究を行っている。高インパクトのスペイン語論文では、地域の労働法、高齢者のための家族健康管理、移民の雇用などのトピックについての社会学者の独自性が失われている [5]。優れた地域的研究の発見・それらへの報奨の付与のためには、高品質の非英語文献に基づいた計量が有用であろう

原則 4　データ収集と分析のプロセスをオープン、透明、かつ単純に保て。

評価のために要求されるデータベースの構成は、明確に表現された規則に従い、研究が終了する前に設定されるべきである。これは、数十年にわたり計量書誌学的評価の方法論を確立してきた学術グループと商業グループに共通の経験である。これらのグループは、査読論文に公表されたプロトコルを参考としてきた。この透明性は精密な検討を可能とした。たとえば、2010 年に、我々のグループのひとつ（ライデン大学の科学技術研究センター(CWTS)）が用いていた重要な指標の技術的性質について公開の討論が行われ、この指標の計算法の改訂に結び付いた [6]。最近参入している商業グループも同様な標準に従うべきである。また、ブラックボックスの評価マシンを受入れるべきではない。

指標が単純であることは、その透明性を増すことであり長所である。しかし、単純化した計量は記録を歪めることもある（原則 7 参照）。評価者は、バランス（研究過程の複雑性に忠実である単純な指標）を得ることに努めなければならない。

原則 5　被評価者がデータと分析過程を確認できるようにすべきである。

データの品質を確かなものにするため、計量書誌学的調査の対象となるすべての研究者が、自分の成果が正確に同定されていることをチェックできるようにすべきである。評価過程の指揮・管理者はすべて、自己確認又は第三者の検査によりデータの正確性を保証すべきである。大学は、その研究情報システムの中にこれを実装することができるだろうし、それは、これらのシステムの提供者の選択の指針であるべきである。正確で高品質なデータの照合・処理には時間と資金を要する。そのための予算を惜しんではならない。

原則 6　分野により発表と引用の慣行は異なることに留意せよ。

ベストプラクティスは、一揃いの指標候補を選び、分野によってその中から選択できるようにすることである。数年前のことだが、欧州のある歴史学者のグループが、その国のピアレビュー評定において比較的低い評点を得たことがあったが、それは、このグループが、Web of Science に収録される雑誌よりもむしろ図書に成果を発表しているためであった。この歴史学者は不運なことに心理学の学科に属していた[歴史学者が心理学の学科に属していたため、雑誌論文によってピアレビュー評

定がなされたという意味だと思われる]。歴史学者や社会科学者は、成果のカウントに際して図書や自国語の論文が含まれることを要求するし、計算科学者は会議論文がカウントされることを要求する。

分野により引用傾向は異なる。トップにランクされる雑誌のインパクトファクターは、数学ではおよそ 3、細胞生物学ではおよそ 30 である。[この差を埋めるための]規格化した指標が必要である。最も頑健な規格化法はパーセンタイルに基づくものであり、各論文は、それが属する分野の被引用数分布中のパーセンタイル位置(たとえばトップ 1%、10%、20%)に従って重み付けされる。非常によく引用される論文 1 件は、パーセンタイル指標に基づくランキングでは、大学の位置を僅かに上げる程度だが、平均被引用数に基づくランキングでは、中位から一挙にトップまで押し上げることがあり得る[7]。

原則 7 個々の研究者の評定は、そのポートフォリオの定性的判定に基づくべきである。

h 指数は、新しい論文がなくても年齢を重ねるほど高くなる。*h* 指数は分野によっても異なる。トップレベルの研究者の場合、生物学では 200、物理学では 100、社会科学では 20-30 程度である[8]。この値は、[*h* 指数の計算に使う]データベースにも依存する。計算科学分野では、Web of Science では *h* 指数が 10 前後であるが、Google Scholar では 20-30 である研究者がいる[9]。研究者の成果物を読んで判定する方が、一つの数字に頼るよりもずっと適切である。多数の研究者を比較する場合でも、個々の専門性、経験、活動及び影響に関するより多くの情報を考慮するやり方が最良である。

原則 8 不適切な具体性や誤った精緻性を避けよ。

科学技術指標は、その概念が曖昧で不確かになりがちであり、また、普遍的には受入れられない強い仮定に立っていることがある。たとえば、被引用数の意味も長らく論争されてきている。従って、ベストプラクティスは、より頑健で複眼的な描像を与えるように複数の指標を用いることである。もし不確かさや誤差が定量化出来るのであれば(たとえばエラーバーの形で)、その情報を公表される指標値とともに示すべきである。それができない場合、指標の作成者は少なくとも誤った精緻性を避けるべきである。たとえば、[Journal Citation Reports では]インパクトファクターを小数点以下 3 桁まで表示して同点の雑誌の出現を避けるようにしている。しかし、被引用数の概念上の曖昧さやランダムな変動性を考慮すれば、このような僅かなインパクトファクターの差によって雑誌を区別する意味はない。誤った精緻性は避けよ。小数点以下 1 桁で十分である。

原則 9 評定と指標のシステム全体への効果を認識せよ。

指標は、それがもたらすインセンティブによってシステムを変化させる。これらの効果を予期しなければならない。このことは、一揃いの指標を用いることが常に望ましいことを意味する。単一の指標は、ゲーム化や目標の取り違えを招く(指標の測定自体が目標になる)。たとえば、1990 年代のオーストラリアでは、機関からの発表論文数に大きく依拠する数式を使って大学の研究への資金配分を行った。大学は査読制雑誌の 1 論文あたりの「価値」を計算することができた。2000 年時点でのその価値は 800 豪ドル(当時のレートで約 480 米ドル)の研究資金に相当した。予想されたように、オーストラリアの研究者が発表する論文数は増加したが、それらは被引用数の低い雑誌に集中し、論文の質の低下を示唆した[10]。

原則 10　指標を定期的に吟味し、改善せよ。

研究の目的と評定の目標は変化し、それに伴って研究システム自体も共進化する。かつて有用であった計量が不適切になり、新しいものが現れる。指標のシステムも見直しが必要であり、適時修正しなければならない。[原則 9 で述べた]単純な数式の影響に気づいて、オーストラリアは 2010 年に、より複雑で質の面を強調した Excellence in Research for Australia イニシアティブを導入した。

次なるステップ

この 10 の原則に従うことにより、研究評価は科学の発展とその社会との関係に重要な役割を果たすことができる。研究計量は、個人の経験だけでは入手・理解することが困難であろう価値ある情報を提供することができる。しかし、この定量的情報を道具から目標に転化させてはならない。

頑健性のある統計を、評価される研究の目標と性質への感受性と組み合わせることによって、最良の決定がなされる。定性的証拠と定量的証拠はそれぞれの方法で客観性を持ち、両者が必要である。科学に関する意思決定は、最高品質のデータに裏付けられた高度の手続きに基づかなければならない。

参考文献

1. Wouters, P. in Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact (eds Cronin, B. & Sugimoto, C.) 47–66 (MIT Press, 2014).
2. Shao, J. & Shen, H. Learned Publ. 24, 95–97 (2011).
3. Seglen, P. O. Br. Med. J. 314, 498–502 (1997).
4. Garfield, E. J. Am. Med. Assoc. 295, 90–93 (2006).
5. López Piñeiro, C. & Hicks, D. Res. Eval. 24, 78–89 (2015).
6. van Raan, A. F. J., van Leeuwen, T. N., Visser, M. S., van Eck, N. J. & Waltman, L. J. Informetrics 4, 431–435 (2010).
7. Waltman, L. et al. J. Am. Soc. Inf. Sci. Technol. 63, 2419–2432 (2012).
8. Hirsch, J. E. Proc. Natl Acad. Sci. USA 102, 16569–16572 (2005).
9. Bar-Ilan, J. Scientometrics 74, 257–271 (2008).
10. Butler, L. Res. Policy 32, 143–155 (2003).

本和訳は Diana Hicks 氏、Nature 誌の許可を得た上で、訳者が独自で行ったものであり、和訳に当たっての原文の解釈に対する全責任を有する。原文では「評価」の概念に含まれる語として“evaluation”、“assessment”、“review”、“judgement”が使われているが、本稿ではそれぞれに対して「評価」、「評定」、「レビュー」、「判定」という訳語を当てた（それらの派生語についても同様）。“metrics”の訳は「計量」に統一した。また、[ ]で示したのは著者による補足である。和訳に際しては、可能な範囲で正確を期しているが、和訳が定まっていない表現も多いことから、より正確な表現については元となる論文を参照願いたい。

この Nature 論文中の 10 項目の原則の部分の日本語訳とともに、ライデン声明の成立の経緯及び解説を含めた記事を、以下に公開している。

研究計量に関するライデン声明について, STI Horizon, Vol.2, No.4, http://doi.org/10.15108/stih.00050