BRITISH JOURNAL OF PSYCHIATRY (2007), 190, 314–318. doi: 10.1192/bjp.bp.106.024919

Fair assessment of the merits of psychiatric research

GRANT LEWISON, GRAHAM THORNICROFT, GEORGE SZMUKLER and MICHELE TANSELLA

Background Use of bibliometric assessments of research quality is growing worldwide. So far, a narrow range of metrics have been applied across the whole of biomedical research. Without specific sets of metrics, appropriate to each sub-field of research, biased assessments of research excellence are possible.

Aims To discuss the measures used to evaluate the merits of psychiatric biomedical research, and to propose a new approach using a multidimensional selection of metrics appropriate to each particular field of medical research.

Method Three steps: (a) a definition of scientific 'domains', (b) translating these into 'filters' to identify publications from bibliometric databases, leading to (c) the creation of standardised measures of merit.

Results We propose using: (a) established metrics such as impact factors and citation indices, (b) new derived measures such as the 'worldscale' score, and (c) new indicators based on journal peer esteem, impact on clinical practice, medical education and health policy.

Conclusions No single index or metric can be used as a fair rating to compare nations, universities, research groups, or individual investigators across biomedical science. Rather, we propose using a multidimensional profile composed of a carefully selected array of such metrics.

Declaration of interest None.

The aims of this paper are to discuss the measures used to evaluate the merits of psychiatric biomedical research, and to propose a new approach using a multidimensional selection of appropriate measures. Such measures can be used, for example, to inform decisions on the allocation of research funds (Lewison *et al*, 1998) within an institution or nationally, or on academic promotions. The choice of which indicators to use is critical. Different single or combined measures can produce entirely different results and implications. We propose that quantitative evaluations of scientific merit in a particular field of biomedical research need a fair set of relevant standardised indicators, chosen according to the type of research being evaluated. The set relevant for each sub-field of research can combine established bibliometric assessments (e.g. impact factors or citation indices); derived measures such as the 'world scale' score discussed below; and new indicators based on journal peer esteem, impact on clinical practice, medical education or health policy.

Our approach comprises three steps, beginning with a clear definition of scientific 'domains' with an agreed set of boundaries, which are then translated into 'filters' to identify relevant publications from bibliometric databases, finally leading to the creation of standardised measures of merit in biomedical research. A domain can be defined at three levels:

(a) the major field, such as biomedical research;

(b) the sub-field, such as psychiatric genetics;

(c) the subject area of a research group, such as the genetics of Alzheimer's disease.

Domain boundaries will not be universally accepted even if they start from agreed definitions (see below). It may be necessary to involve several experts in order to reach a consensus. Focusing at the second level (sub-fields), we illustrate each of these three steps using as examples the sub-fields of psychiatric genetics and health services research in mental health, which differ in representing respectively the more basic and the more applied ends of the research spectrum (Dawson, 1997; Horig & Pullman, 2004).

METHOD

Defining the scientific domains of interest

We use the following definition of psychiatric genetics:

> 'The role of genes in mental disorders, investigated through family linkage or association studies (sometimes involving polymorphisms in candidate genes). Effects may be observed through psychopathology, psychopharmacology, personality, cognitive function or behavioural variation' (Kendler, 2005).

An understanding of health services research in relation to mental health also depends upon its definition. This has been given, for example, by the Medical Research Council in the UK (Medical Research Council, 2002), AcademyHealth in the USA (AcademyHealth, 2004) and the Health Services Research Hedges team (Wilczynski *et al*, 2004). From reviewing these we propose that health services research be defined as the multidisciplinary field of scientific investigation that:

(a) describes healthcare needs, variations in access to services, and patterns and quality of healthcare provision;

(b) evaluates the costs and outcomes of healthcare interventions for individuals to promote health, prevent or treat disease or improve rehabilitation;

(c) determines ways of organising and delivering care for populations;

(d) develops methods of disseminating evidence-based practice;

(e) investigates the broader consequences of healthcare interventions, including acceptability, effects on carers and families, and the differential impact of interventions on subgroups of patients.

Health services research in mental health is therefore considered here as the application of this definition to mental disorders, their treatments and related services.

Developing filters to identify scientific publications in a specified domain

Filters have been developed over the past decade to identify in bibliometric databases publications that are relevant to the 'cause, course, diagnosis or treatment' of specific health problems (Haynes *et al*, 2005) to 'aid clinicians, researchers and policy-makers harness high quality and relevant information'. For particular sub-fields, specialist journals have traditionally been used to identify relevant publications, but need to be supplemented with title words, often in combination (Lewison, 1996), because for many biomedical sub-fields two-thirds or more of the papers will be in 'general' journals. Such a filter can achieve both a specificity (or precision) and a sensitivity (or recall) above 90%, and calibration methods have been described (Lewison, 1996).

In relation to our illustrative sub-fields, the filter developed for psychiatric genetics selected papers from the Science Citation Index if they were within both the sub-fields of genetics and mental health. Of the resulting scientific papers identified, 93% were relevant; this rose to 99% when those on 'suicide genes' (which are not related to mental health) were removed. The mental health services research filter was also based on the intersection of two separate filters (health services research and mental health), but that for health services research was much harder to define (Wilczynski *et al*, 2004). Although its specificity was as high as 0.93, the sensitivity only reached 0.59, as it proved difficult to list all the combinations of title words on many relevant papers. In principle, sensitivity can be improved by incorporation of additional title words or journals taken from false-negative papers from relevant departments. However, this may be at the expense of specificity if too many terms are included in the filter.

RESULTS

Established measures of the merits of biomedical psychiatric research

Research evaluation is concerned both with the volume of output and its quality. Regarding volume, the number of identified research papers can be used at a global, national or institutional level to consider whether the amount of research is commensurate with the associated disease burden (Lewison, 2005). There may be an international imbalance, as was shown by the Global Forum for Health Research, for example for AIDS (de Francisco, 2004). At the national level, the correlation between numbers of papers and global burden of disease was found to be good for deaths from gastric cancer (r^2=0.90; Lewison *et al*, 2001) but very poor for those from lung cancer (r^2=0.04; Rippon *et al*, 2005). At the institutional level, publication counts can be compared with inputs of money and personnel.

In relation to the quality of publications, the central problem is that most evaluations of scientific merit are limited to the number of citations of papers by other papers (as recorded in the Science Citation Index or the Social Science Citation Index), or to analyses of journal impact factors (Tsafrir & Reis, 1990; Seglen, 1997). These may be more appropriate for the basic science sub-fields such as psychiatric genetics, where citations are numerous, but may give a distorted view of applied clinical sub-fields, and so may prejudice applications for competitive funding.

The 'world scale' for assessing domains of research

To complement these two established measures, we propose a new scale assessing the relative scientific merit of a country or of an institution. This new scale is derived from citation scores or from journal impact factors. It is based on the concept of World-scale, an idea borrowed from the oil tanker charter market, in which the output from an entity (a country, an institution or an individual) is compared with that of the world at different levels of excellence. One might ask, for example, what percentage of the output of such an entity receives a citation score sufficient to place it in the top 10% of the world production in that particular domain: if it is more than 10%, this indicates a superior performance, and if it is less, then it is not so meritorious on the selected criterion. Similar calculations can be made at other centiles (e.g. 5%, 20%) and an average value determined, or one weighted to reflect the greater importance of performing well at the top levels. The 5-year citation window is used because it strikes a balance between the need to allow time for the papers to be properly judged by the scientific community and immediacy. World scale values could also be based on shorter (or longer) time windows.

Table 1 shows world scale values for UK and US papers in the selected sub-fields, using citation scores as the source data. For example, the USA has 25 papers from 706 in psychiatric genetics that received 112 or more citations in the given period, or

Table 1 Citation scores for papers in psychiatric genetics and mental health services research for 1996–98, cited in the year of publication and four subsequent years

				US papers		UK papers	
Centile (%)	Cites[1]	World[2]	Actual centile (%)	*n*	World scale value	*n*	World scale value
Psychiatric genetics[3]							
2	112	30	2.03	25	175	5	101
5	65	76	5.13	56	155	12	96
10	42	149	10.06	107	151	27	110
20	24	309	20.86	206	140	60	118
All		1481		706	155	244	106
Mental health services research							
2	35	45	2.07	40	167	4	52
5	21	123	5.65	92	140	19	91
10	15	226	10.38	169	140	33	86
20	9	471	21.64	321	128	76	95
All		2177		1161	144	369	81

1. Number of citations over 5 years needed for a paper to be in the respective centile.
2. Actual number of world papers that qualify in this way.
3. Psychiatric genetics data are taken from the Science Citation Index (CD–ROM version).
4. Mental health services research data are taken from the Science Citation Index and the Social Sciences Citation Index (CD–ROM version).

https://doi.org/10.1192/bjp.bp.106.024919 Published online by Cambridge University Press

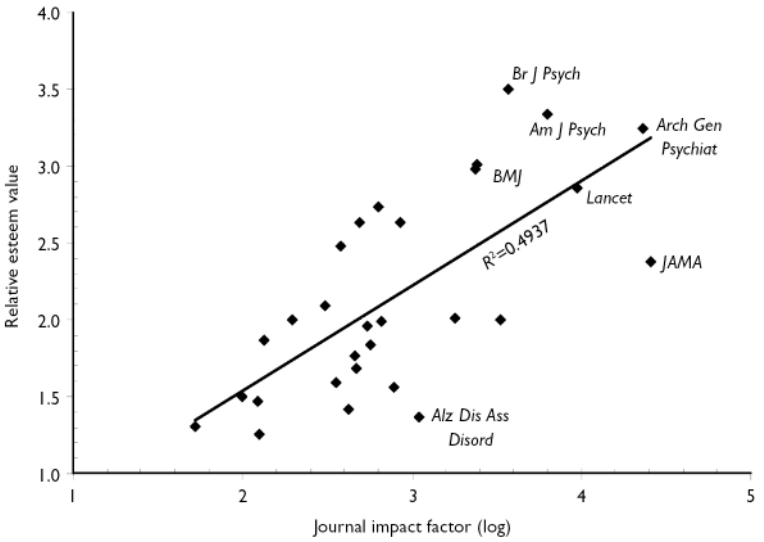

Fig. 1 Comparison of the relative esteem value given to 29 leading journals used for mental health services research with their impact factors (a log scale is used, which relates more closely to the subjective view of researchers than does the crude impact factor). *Alz Dis Ass Disord*, *Alzheimer Disease and Associated Disorders*; *Am J Psych*, *American Journal of Psychiatry*; *Arch Gen Psychiat*, *Archives of General Psychiatry*; *Br J Psych*, *British Journal of Psychiatry*.

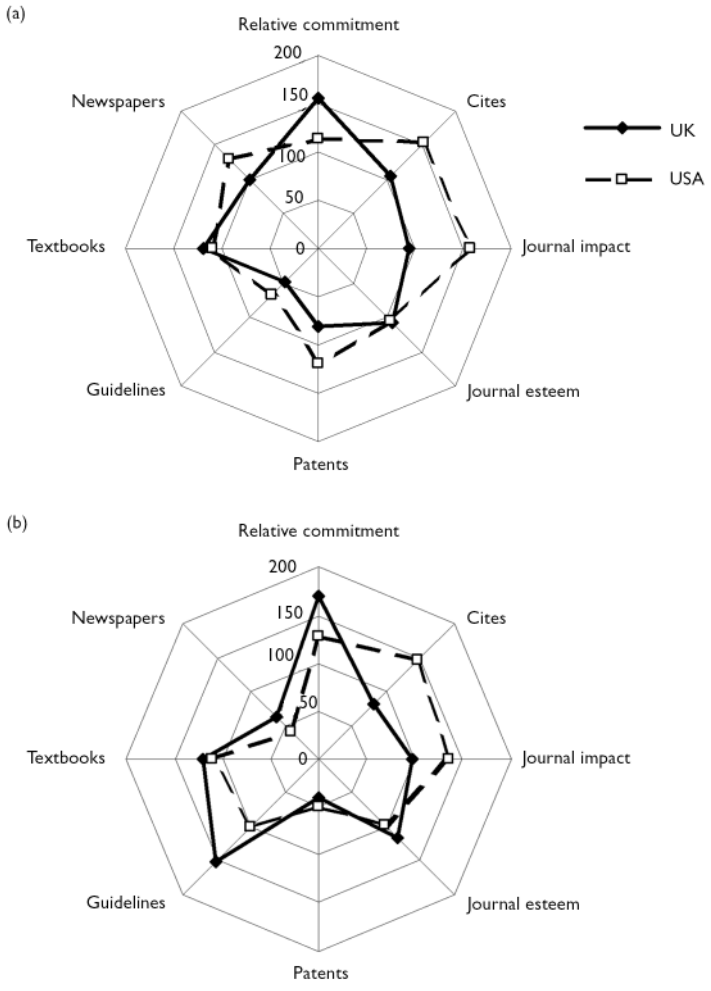

Fig. 2 Kite diagrams for (a) psychiatric genetics and (b) mental health services research. Real data are used for USA and UK relative commitment, citations, journal impact and journal esteem; dummy data are used for citations on patents, on clinical guidelines, in textbooks and in newspapers.

3.54% compared with the world norm of 2.03%, so its world scale value at the 2% centile was $(3.54/2.03) \times 100 = 175$. We can see that the USA has a superior performance over the whole range of centiles in both sub-fields; the UK is slightly better than average in psychiatric genetics, but below average (especially at the higher centiles) in mental health services research, probably because work in this area tends to be more specific to a country, with UK experience less relevant to USA researchers, who published over 53% of all the papers (a reason for distrusting citation scores alone in such a domain).

World scale values can also be calculated from journal impact factors, and the US values are 157 for psychiatric genetics and 136 for mental health services research. In comparison, the UK psychiatric genetics score is 94 and the mental health services research score is 98, which reverses the trend seen above in world scale values based upon citation counts.

Relative esteem value

A further measure that can complement impact factors and citation indices in assessing scientific merit is the 'relative esteem value' of journals (Lewison, 2002; Jones *et al*, 2004). This is determined from written questionnaires to researchers in a sub-domain, which invite them to rate journals on a scale from 'excellent' to 'decidedly secondary'. For the more basic sub-fields there is a reasonable correlation with journal impact factors (about 0.6), but in more applied clinical specialties the correlation coefficient may drop to zero (Lewison, 2002), with some highly cited journals being of lower subjective esteem for communication of research results than some less frequently cited journal (Lee *et al*, 2002). Figure 1 shows the relationship between relative esteem value and impact factor for 29 leading journals in mental health services research.

World scale values can also be calculated using relative esteem values, as previously described (Lewison, 2004). Comparing the USA and UK for mental health services research using a world scale based upon such values, rated by 88 international researchers in the field, we found scores of 98 for the USA and 116 for the UK, again giving quite different results from the world scales from citation counts (144 for the USA and 81 for the UK). The relative standing of the scientific output

https://doi.org/10.1192/bjp.bp.106.024919 Published online by Cambridge University Press

from different countries (or institutions) in particular sub-fields can therefore be highly dependent upon the assessment measures used.

Multidimensional assessment of scientific merit

No single measure alone can therefore provide a stable and rounded assessment of merit within a scientific sub-field. Rather, we propose that a range of indicators be used for a full appreciation of the value of research in any particular sub-field. These may include a combination of some of the following: impact factors; citation indices; world scale values; or relative esteem values, along with counts of patents that cite references within the sub-field (mainly for basic research); citations in clinical guidelines (mainly for clinical studies) (Grant *et al*, 2000); citations in journals actually read routinely by clinicians; citations in newspapers that are read by policy makers, healthcare professionals, researchers and the general public; citations in governmental and international policy documents (Lewison, 2004); citations in relevant international standards; citations in textbooks, which can indicate an impact on medical education (Lewison, 2004); and presence of researchers on journal editorial boards.

The comparative quality of research in a country, for example, can be shown in a multidimensional graphical display such as a kite diagram. Figure 2 shows such kite diagrams for the USA and the UK in the two sub-fields across eight indicators of scientific merit. These profiles use a further measure, the 'relative commitment' score: this measures the amount of effort a country devotes to a scientific sub-field, compared with its overall biomedical research portfolio, relative to the world average (world scale). For example, the UK publishes about 17% of world papers in mental health services research, whereas its presence in the world biomedical literature is only 10%, so its relative commitment is $17/10 \times 100 = 170$ on the world scale index in this sub-field. Figure 2 therefore shows that the UK performs well in both sub-fields, but the next two indicators, moving clockwise, namely citations and journal impact, show that UK output has less impact than US publications in both sub-fields. Figure 2 also suggests that new indicators can be developed that quantify other important dimensions of research impact, such as informing clinical practice (Perneger, 2004), or enhanced patient safety (Agoritsas *et al*, 2005), for example as assessed through scientific paper citations in clinical guidelines or protocols. The other four indicators (patents, guidelines, and textbook and newspaper citations) are illustrated in Fig. 2 with dummy values, and can be determined in practice for such an evaluation to be complete. Such diagrams can also be used to compare institutions.

DISCUSSION

The range of assessment methods used in specific sub-fields of research should be appropriate to each case. In our examples, psychiatric genetics and mental health services research, it may be more important for the latter than for the former to influence health policy, treatment guidelines and clinical practice (Institute of Medicine, 2001). It is therefore reasonable to develop a range of such measures of health services research impact, but it would be unreasonable to apply all of them to assess psychiatric genetics. By the same token, if measures developed for the more basic biomedical sciences are used uncritically in the applied sciences, the latter may apparently perform poorly, and consequently suffer in terms of resource allocation (Dash *et al*, 2003).

However, more statistics do not mean better statistics. The key questions remain: who should make the choices between different measures of scientific merit; who should decide how these criteria are weighted; and with what agenda? We propose that the use of these measurements is best done within the context of the peer review process, as this is the strongest method so far devised to assure an overall appraisal of scientific merit. No single index or metric can be used as a fair rating to compare nations, universities, research groups or individual investigators between different sub-fields of science (Goldberg & Mann, 2006). Rather we propose that research oversight and peer review procedures refer not to any single measure of research quality (as is often the case at present), but refer simultaneously to a multidimensional profile, composed of a carefully selected array of such metrics (Martin, 1996), to construct a balanced and fair assessment of the merits of psychiatric research.

GRANT LEWISON, BA, PhD, Evaluametrics Ltd., Kew, Richmond, Surrey and School of Library, Archive and Information Studies, University College London; GRAHAM THORNICROFT, FRCPsych, FMedSci, Health Services and Population Research Department, Institute of Psychiatry, King's College London, UK; GEORGE SZMUKLER, FRCPsych, Institute of Psychiatry, King's College London, UK; MICHELE TANSELLA, MD, Department of Medicine and Public Health, Section of Psychiatry and Clinical Psychology, University of Verona, Italy

Correspondence: Professor Graham Thornicroft, Health Service and Population Research Department, King's College London, De Crespigny Park, London SE5 8AF, UK. Tel: +44(0)20 7848 0735; fax: +44(0)20 7277 1462; email: g.thornicroft@iop.kcl.ac.uk

(First received 30 March 2006, final revision 25 August 2006, accepted 27 October 2006)

REFERENCES

AcademyHealth (2004) *Glossary of Terms Commonly Used in Health Care*. AcademyHealth.

Agoritsas, T., Bovier, P. A. & Perneger, T. V. (2005) Patient reports of undesirable events during hospitalization. *Journal of General Internal Medicine*, **20**, 922–928.

Dash, P., Gowman, N. & Traynor, M. (2003) Increasing the impact of health services research. *BMJ*, **327**, 1339–1341.

Dawson, S. (1997) Inhabiting different worlds: how can research relate to practice? *Quality of Health Care*, **6**, 177–178.

De Francisco, A. (2004) Measuring the 10/90 gap: comparing disease burden with investment in health research. In *The 10/90 Report on Health Research 2003–2004*, pp. 122–125. Global Forum for Health Research.

Goldberg, D. & Mann, A. (2006) How should financial support for research be distributed to universities? The Research Assessment Exercise (RAE) in England and Wales. *Epidemiologiae Psichiatria Sociale*, **15**, 104–108.

Grant, J., Cottrell, R., Cluzeau, F., et al (2000) Evaluating payback on biomedical research from papers cited in clinical guidelines: applied bibliometric study. *BMJ*, **320**, 1107–1111.

Haynes, R. B., McKibbon, K. A., Wilczynski, N. L., et al (2005) Optimal search strategies for retrieving scientifically strong studies of treatment from Medline: analytical survey. *BMJ*, **330**, 1179–1184.

Horig, H. & Pullman, W. (2004) From bench to clinic and back: perspective on the 1st IQPC Translational Research conference. *Journal of Translational Medicine*, **2**, 44.

Institute of Medicine (2001) *Crossing the Quality Chasm: A New Health System for the 21st Century*. Institute of Medicine.

https://doi.org/10.1192/bjp.bp.106.024919 Published online by Cambridge University Press

Jones, T., Hanney, S., Buxton, M., *et al* (2004) What British psychiatrists read: questionnaire survey of journal usage. *British Journal of Psychiatry*, **185**, 251–257.

Kendler, K. S. (2005) Psychiatric genetics: a methodologic critique. *American Journal of Psychiatry*, **162**, 3–11.

Lee, K. P., Schotland, M., Bacchetti, P., *et al* (2002) Association of journal quality indicators with methodological quality of clinical research articles. *JAMA*, **287**, 2805–2808.

Lewison, G. (1996) The definition of biomedical research subfields with title keywords and application to the analysis of research outputs. *Research Evaluation*, **6**, 25–36.

Lewison, G. (2002) Researchers' and users' perceptions of the relative standing of biomedical papers in different journals. *Scientometrics*, **53**, 229–240.

Lewison, G. (2004) Citations to papers from other documents: evaluation of the practical effects of biomedical research. In *Handbook of Quantitative Science and Technology Research* (eds H. Moed, W. Glaenzel & U. Schmoch), pp. 457–472. Kluwer.

Lewison, G. (2005) Biomedical research and the regional burden of disease. In *Proceedings of the Tenth Conference of the International Society for Scientometrics and Informetrics* (eds P. Ingwersen & B. Larsen), pp. 585–594. Karolinska University Press.

Lewison, G., Cottrell, R. & Dixon, D. (1998) Bibliometric indicators to assist the peer review process in grant decisions. *Research Evaluation*, **8**, 47–52.

Lewison, G., Grant, J. & Jansen, P. (2001) International gastroenterology research: subject areas, impact, and funding. *Gut*, **49**, 295–302.

Martin, B. (1996) The use of multiple indicators in the assessment of basic research. *Scientometrics*, **36**, 343–362.

Medical Research Council (2002) Cluster randomised trials. MRC.

Perneger, T. V. (2004) Relation between online hit counts and subsequent citations: prospective study of research papers in the BMJ. *BMJ*, **329**, 546–547.

Rippon, I., Lewison, G. & Partridge, M. R. (2005) Research outputs in respiratory medicine. *Thorax*, **60**, 63–67.

Seglen, P. O. (1997) Why the impact factor of journals should not be used for evaluating research. *BMJ*, **314**, 498–502.

Tsafrir, J. S. & Reis, T. (1990) Using the citation index to assess performance. *BMJ*, **301**, 1333–1334.

Wilczynski, N. L., Haynes, R. B., Lavis, J. N., *et al* (2004) Optimal search strategies for detecting health services research studies in MEDLINE. *Canadian Medical Association Journal*, **171**, 1179–1185.

https://doi.org/10.1192/bjp.bp.106.024919 Published online by Cambridge University Press