*Jointly published by Elsevier Science Ltd, Oxford*
*and Akadémiai Kiadó, Budapest*

*Scientometrics,*
*Vol. 41, Nos 1–2 (1998) 5–16*

# NEW BIBLIOMETRIC TECHNIQUES FOR THE EVALUATION OF MEDICAL SCHOOLS

G. LEWISON

*Unit for Policy Research In Science and Medicine (PRISM),*
*The Wellcome Trust, 210 Euston Road, London NW1 2BE (England)*

(Received August 30, 1997)

Bibliometrics have been used in novel ways to assist with the evaluation of two medical schools, one in England and one in Sweden. The first evaluation was intended to allow the relative strengths in 26 subfields of five component campuses to be estimated. Selective filters for each subfield were defined, many of them with the help of the school's research staff, so that relevant papers could be retrieved from a database on the basis of their title keywords and specialist journals. The campus outputs were then analysed by the research level of the journals (clinical/basic) and their influence. In the second evaluation, nine different indicators of research output were produced so that the school could be compared with four others in Scandinavia. The indicators included measures of output, co-authorship, journal esteem and citations by papers and by patents.

## Introduction

Although bibliometrics are now widely seen as having an important place in the evaluation of research (*Narin* and *Hamilton* 1996; *Daniel* and *Fisch* 1990; *van Raan* 1993) they are still viewed with suspicion by some of those being evaluated (*Collins* 1991). It is desirable, therefore, that they should cover many different aspects of research outputs (*Martin* 1996) and that the evaluees can play a part in helping to create appropriate methodology. The latter process can make a big difference to the way the results are accepted. This paper describes the methods used in two quite different evaluations in which bibliometrics played a major role, and the lessons learned which may have wider applicability. The evaluations concerned medical schools, which aim to cover almost all areas of biomedical research as a pre-requisite to effective medical education and to the local provision of clinical care to a high standard in most specialties.

*0138–9130/98/US $ 15.00*

The first evaluation was of the Imperial College Medical School (ICMS) in London. This school is being formed from five different campuses:
Charing Cross and Westminster Medical School;
Imperial College Main Site including the Silwood Park Field Station;
National Heart and Lung Institute and Brompton Hospital;
Royal Postgraduate Medical School/Hammersmith Hospital;
St Mary's Hospital and Medical School;
each of which has its own traditions and specialities. It was required to provide an overview of the strengths and weaknesses of each one in terms of their research capability in some 26 subfields. Research capability was taken to mean the output of papers in the serial literature for years 1991-94 from the campuses (defined by their individual postcodes) and by staff currently in post who had joined since 1991 from elsewhere in the UK, or who held a joint appointment with another hospital. It was considered impractical to subtract out papers from staff who had recently left because many of these papers would be co-authored with current staff.

Since the main purpose of the evaluation was to examine outputs of papers in individual subfields, these had to be defined in the form of "filters" that would selectively retrieve relevant papers from PRISM's Research Outputs Database (*Jeschin* et al. 1995). These filters were created by an iterative process, mostly with the assistance of senior research staff from the ICMS. This procedure is described in detail elsewhere (*Lewison* 1996): it involved visits by the researchers to the PRISM office to help define the filter and the "marking" of lists of titles and journal names of sets of papers retrieved by the filter to give practical expression to the experts' views of what each subfield contained. Within each subfield, the study was intended to show not only the position of each of the five campuses, but also that of the Imperial College Medical School as a whole, within the context of the UK research output. However there was considerable concern that the filters would be unable to produce "clean" lists of papers from each campus, and in particular that they would omit the basic research papers without distinctive title keywords published in prestigious journals, and so would not do justice to some groups of researchers.

The second evaluation was of the Göteborg University Faculty of Medicine, which was undertaken during 1996 by an independent international panel chaired by Dr John Bienenstock, Dean and Vice President of the Medical School at McMaster University in Hamilton, Ontario (*Bienenstock* et al. 1996). The panel asked the university to commission bibliometric studies from the University of Leiden in the Netherlands and from CHI Research, Inc. in the USA in order to inform their deliberations. The studies were intended to provide a quantitative measure of comparison of Göteborg with the medical schools in four other Scandinavian cities of similar biomedical research output (Copenhagen, Helsinki, Lund and Oslo), and to reveal which were the strong and which were the weaker research subfields for Göteborg.

Because of the need to protect the panel's independence, it was harder to involve the faculty members in these studies, but they were consulted about the respective weights to be given to papers in particular groups of journals and about the categorization of the journals into three groups. The citations to the papers were analysed and the results presented using a new method, which took account of citations to all the papers in a cohort and not just of the mean, itself a statistic of dubious value (*Anderson* 1989). Here the concerns of the researchers for a fair evaluation were met by the use of many different indicators and, for the appraisal of which subfields were strong and which weak, by the use of two independent methods.

## Methods

For the Imperial College study, the outputs were displayed by means of carpet plots of numbers of papers in each of the 4×4=16 cells of a matrix, whose axes were journal research level (clinical *vs.* basic) and journal quality (ordinary to excellent). The research level of almost all the journals processed for the *Science Citation Index*, on which the Research Outputs Database is currently based, has been assigned by CHI Research Inc. (*Narin* et al. 1976) on the basis of expert opinion and journal-to-journal citation patterns into four categories (Table 1).

Table 1
Categories of research level for journals

| Research level | Description |
|---|---|
| 1 | clinical observation |
| 2 | clinical mix |
| 3 | clinical investigation |
| 4 | basic research |

The assignment of journals into quality categories is not invariant but depends upon the subfield, the more popular ones with heavily cited journals having tougher entry standards for the higher weighted categories. The basic scheme adopted was to put the top 10% of the core set of journals in each subfield into the "excellent" category with weighting, W = 4; the next 20% into the "very good" category with W = 3; the next

30% into the "good" category with W = 2. The remaining 40% of core journals were regarded as "ordinary" with W = 1. The ranking was made on the basis of mean citation scores of 1990 papers cited from 1990 to 1994 (mean $C_{0-4}$ values), which are tabulated by the Institute for Scientific Information in the "Journal Expected Citation Rates" file. For new journals not published in 1990, mean values of $C_{0-3}$ and $C_{0-2}$ were determined for 1991 and 1992 papers respectively cited through 1994, as available, and extrapolated to a mean $C_{0-4}$ value on the basis of a simple formula. Once the core set of journals for the subfield, normally consisting of the most frequently used 10% of the total set and accounting for some two thirds of all publications retained by the filter, had been selected and allocated to the four weighting categories, the critical mean $C_{0-4}$ values needed for a journal to be accepted in each class were then applied to the remaining 90% of journals so as to give them each a W value. In this way, the weighting was based on the journals well-known to researchers in the subfield and not biassed by the presence of literally many hundred journals that were rarely used by the subfield's researchers for their published work.

In order to satisfy the ICMS researchers' concern that the filters would omit some important papers, the lists of papers in each subfield were "cleaned up" by a two-stage process. First, lists of all ICMS papers retrieved by each subfield filter were prepared and circulated to appropriate experts, who were invited to mark any papers considered irrelevant to the subfield (false positives) and suggest in which other subfields they might lie. Adjustments were then made to the attributions of the papers so marked. The second step was to prepare lists of papers for each of the five campuses that had not been attributed to any of the subfields, and invite a representative from each one to indicate to which subfield(s) they should be allocated. In this way, false negatives would be corrected.

For the Göteborg study, the first step was to identify all the papers (articles, notes and reviews) from medical school addresses in each of the five cities over a 10-year period, 1986 to 1995, within the *Science Citation Index*. These were then used to determine the partial indicators shown in Table 2.

The individual indicators were presented in detail as an *Annex* to the evaluation report and brought together in a single table to show the ranking of the five cities on each of the nine partial indicators.

Table 2
Indicators of relative performance used for Göteborg evaluation

| Indicator | Comment |
|---|---|
| Mean annual output | Graphs were drawn showing running 3-year means |
| Increase in output | Between 1986-88 and 1993-95, a 7-year period |
| Amount of international co-authorship | Shows how open the universities are to working with foreign scientists |
| Increase of international co-authorship | Should take account of some countries recently joining the EU |
| Percentage of reviews | Shows how many senior scientists have a high reputation |
| Journal esteem factor | Based on weighting papers in eight "A" journals @ 5.5 and ones in 291 "B" journals @ 2.5. |
| Citations to 1986 papers | The percentage of papers from the city that are cited in the top 5% of the whole group of papers from all five cities |
| Citations to 1991 papers | The citation standard rose for the whole group between the two years by about 20% |
| Citations by US patents | Shows utility for industrial innovation. Based on norm of the whole group of papers from 5 cities. |

## Results

The purpose of this paper is to describe the methodology rather than to give definitive results, so only a few sample results will be given in order to illustrate the methods used and point out their advantages and limitations. Figure 1 shows the 4×4 carpet plot of outputs broken down by research level and journal weighting for one subfield (cardiovascular research) and one of the five ICMS campuses, for papers retrieved directly by the filter (gross output). The filter had an overall precision, p, of 0.96 and an overall rate of recall, r, of 0.88, so the ratio of the actual number of cardiovascular papers to those found would have been $p/r = 1.10$. Journals are given a W value of 4 if their mean $C_{0-4}$ value>17.6; W=3 if mean $C_{0-4}>10.9$; W=2 if $C_{0-4}>6.0$; and W=1 otherwise.

For Göteborg and the other Scandinavian cities, the first four indicators are straightforward and require no comment. The fifth indicator, the percentage of reviews, was regarded as something of an experiment. It is very simple to calculate and Fig. 2 shows that there is a strong and positive correlation between a nation's percentage of reviews and its citation performance. Reviews are frequently written by invitation and so they might be regarded as an appropriate indicator of the esteem in which a country's or a city's researchers are held by the scientific community. However some countries depart from this general trend: Russia and the UK write more reviews than would be expected from the recent impact of their publications; conversely China and Japan write fewer.

The journal esteem factor was based on the attribution of weights greater than unity to papers from each of the five cities that were published either in a small set of eight prestigious "A" journals (*Cell, EMBO J, J. Clin. Investigation, Lancet, Nature, New Engl. J. Med, Proc. Nat. Acad. Sci. USA, Science*) or in a much larger set of good or "B" journals. These were defined as the top 10% non-review journals in each subfield on the basis of impact factors (which were recalculated by the University of Leiden), plus any review journals in the subfield with impact factors above that of the lowest non-review journal retained, plus any other journals with IF>3.0. The total "B" set numbered 291 journals of which 67 were review journals. Altogether, 1% of the papers from the five cities were in "A" journals and 19% in "B" journals.

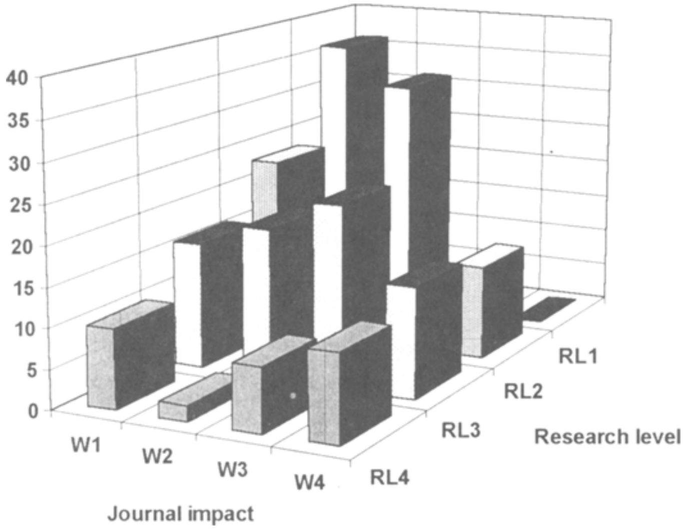

Fig. 1. Carpet plot of papers from an ICMS campus in cardiovascular research, 1991–1994

The weights to be used were obtained in part from a survey of 19 Göteborg professors, who on average voted 5 for papers in "A" journals and 2.5 for papers in "B" journals, and in part from a survey (*Lewison* 1995) of some 40 scientific administrators in the UK who voted 6 for papers in "excellent" journals (corresponding to "A") and 2.5 for papers in "good" journals (corresponding to "B"). Mean values of 5.5 for "A" and 2.5 for "B" were taken and the results are shown in Fig. 3 as three-year moving averages of the "journal esteem factor", calculated as the weighted sum of the numbers of papers from each city in the different journals divided by the simple sum. Helsinki has consistently published its papers in the "best" journals, but Göteborg lies second.

Figure 4 shows the distribution of citations over a five-year period starting with the publication year for all the publications from the five cities for 1986 and 1991. The ordinate is the number of citations, $C_{0-4}$, needed for a paper to be in the centile given by the abscissa on a log scale. There appears to have been an "inflation" of citation numbers by about 20% between the two years. For the papers in 1991 from each city, the numbers with $C_{0-4}$ values 79 and above and so high enough to put them into the first centile of the whole cohort; 56 and above and so high enough to put them into the second centile, etc., were determined. For the Göteborg papers, 1.52% had 56 citations or more, so the relative centile presence, or RCP, was 1.52%/2% = 0.76 at the second centile, but for the Helsinki papers, 3.04% had 56 citations or more, so their RCP was 3.04%/2% = 1.52. The values of RCP at each of the centiles, 1%, 2%, 5%, 10%, 20% and 50% (median) were plotted to give the curves of Fig. 5. This shows how the papers from Helsinki were the most cited at all centiles above the 30th, but the papers from the other cities are similar in their citation performance.

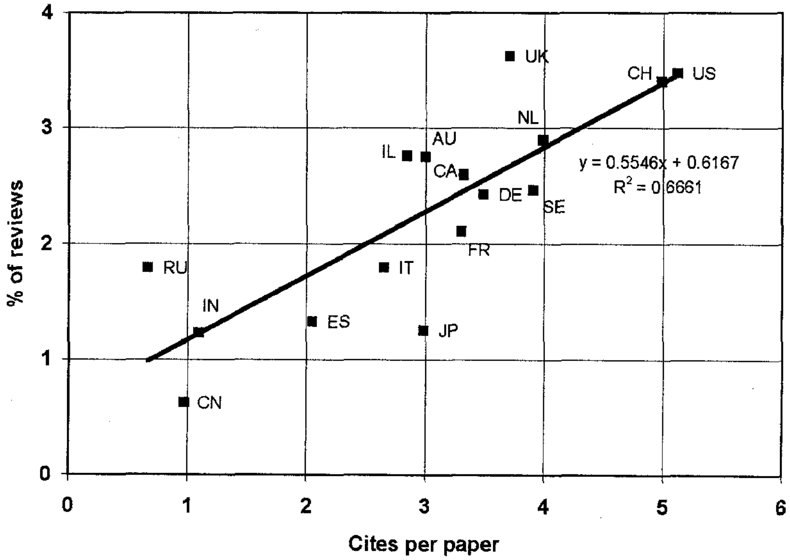

Fig. 2. Percent of reviews vs. cites per paper, 1989–1993, for all fields of science for 16 countries

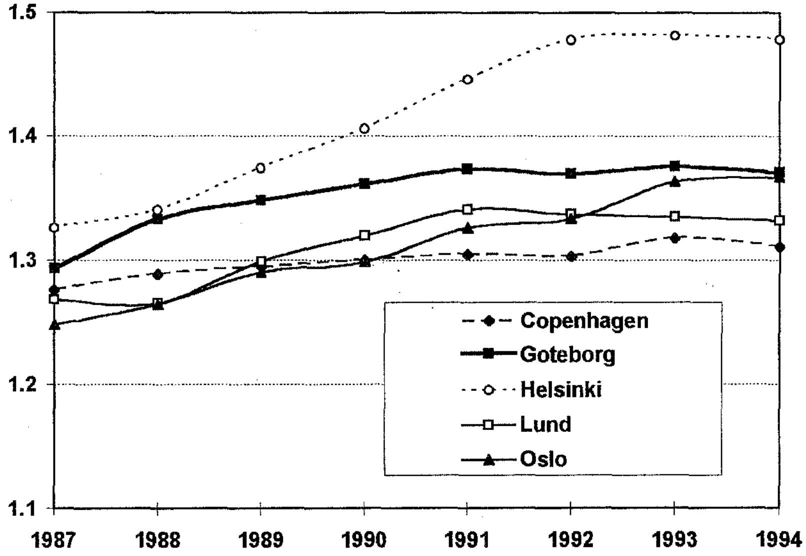

Fig. 3. Journal esteem factor for medical research papers from five Scandinavian cities, 1987–1994, 3-year moving averages

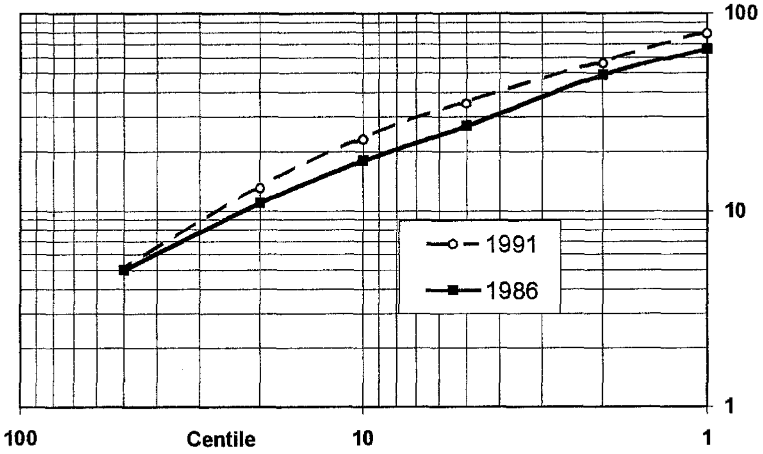

Fig. 4. Distribution of citations to medical research papers published in 1986 and 1991 from group of five Scandinavian cities, 5-year citing window

Citation of papers by patents is becoming increasingly frequent and especially in advanced areas of biomedicine such as human genetics (*Anderson* et al. 1996) but it is still far less common than citations by other papers. For United States Patent and Trademark Office (USPTO) patents up to the end of 1995, the chance that a Scandinavian biomedical paper from any one of the five cities is cited by a patent is shown in Fig. 6. It rises from effectively zero for 1994 publications to 1% for 1991 papers and about 2.4% for 1987 ones.

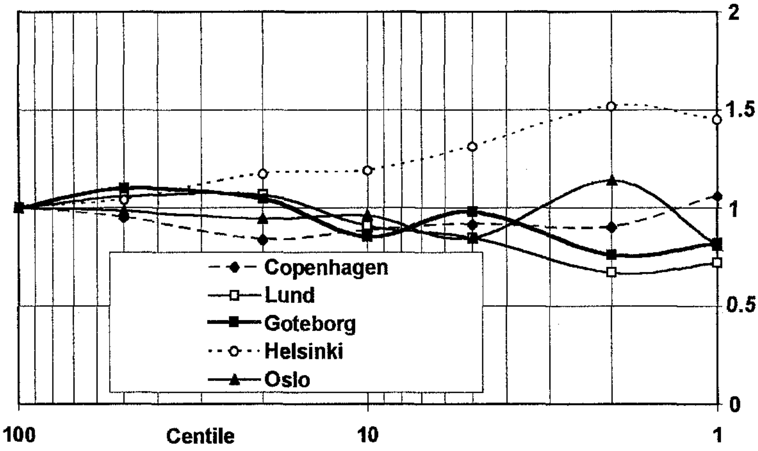

Fig. 5. Relative centile presence for 1991 medical research papers from each of 5 Scandinavian cities cited through 1995, compared to overall norm for the five

The average number of cites per cited paper rises also, but only from 1.0 to 1.6 over the same period. Relative to these average citation rates, the actual rates for papers from the individual cities varied, with the papers from Lund the most useful in underpinning commercially valuable technology, as embodied in patents, followed by those from Göteborg and Helsinki.

Finally, Fig. 7 shows the relative strength of Göteborg in subfields for which two definition schemes were available, one based only on specialist journals (CHIrank) and one based also on title keywords, described above (PRISMrank). Some 19 subfields could be investigated. The CHIrank shows the position of Göteborg relative to the other four cities as a group for 35 subfields, the highest, cardiovascular research = CARDI, being scored 100 and the lowest, radiology & nuclear medicine, zero. The PRISMrank shows the position of Göteborg relative to Sweden as a whole for 21 subfields, again the one with its greatest relative presence, primary healthcare, being scored 100 and the lowest, structural biology, zero.

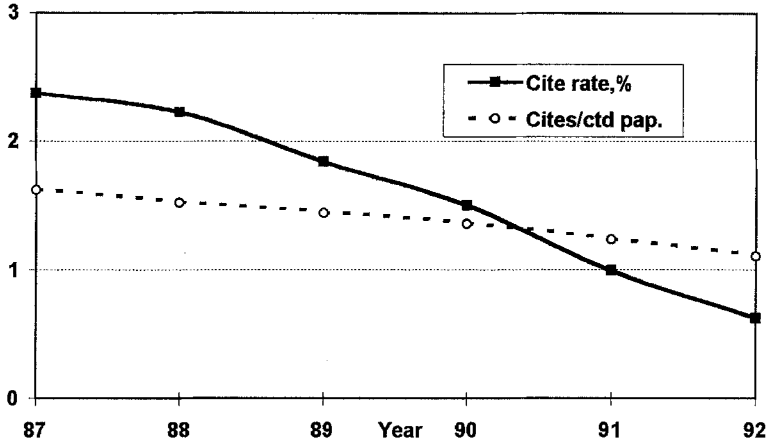

Fig. 6. Rate of citation to medical research papers from group of five Scandinavian cities by USPTO patents up to 1995 and cites per cited paper: 3-year moving averages

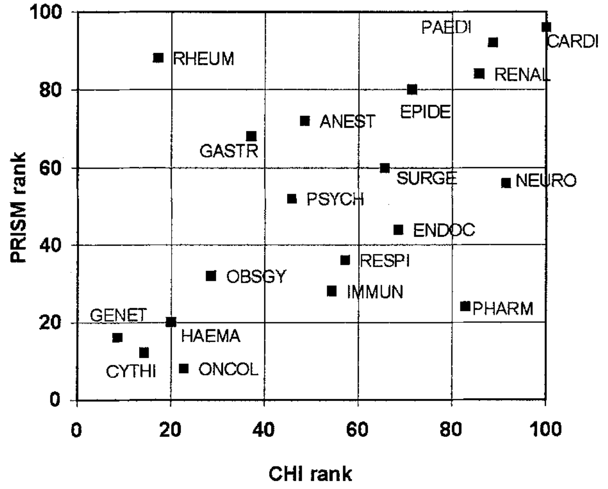

Fig. 7. Comparison of ranking of Göteborg in different subfields by two methods

## Discussion

The main lesson from the first evaluation, of the Imperial College Medical School, was that the researchers were rather suspicious of the bibliometric approach and much time and patience was needed in order to secure their co-operation or at least their acquiescence. Many researchers doubted that all their papers had been counted, but on investigation nearly all of the missing ones turned out to be items other than articles, notes and reviews, or in journals not covered by the *Science Citation Index* or *Social Sciences Citation Index*, or not in journals at all.

The process of subfield definition, although it involved a fair amount of "homework" in which senior researchers marked papers as relevant, marginal, or irrelevant, was very helpful in this respect. The correction of the lists of papers, first to allow for new arrivals and joint appointments, and secondly to correct for false positives and then for false negatives, further improved matters, although at the expense of a lot of work by both PRISM staff and those being evaluated. It turned out that some subfield filters were deemed to be generating a lot of papers outwith their subfield. In particular, the precision of developmental biology, epidemiology, haematology and primary care was poor, perhaps because of differing concepts of what the subfields should comprise. However the second step, in which "unattributed" papers were allocated to subfields by campus representatives, worked very well, and almost 90% of these papers were so attributed.

The lesson from the second evaluation, of the Göteborg University Medical Faculty, was that different indicators can give very different pictures of the research capability of an institution and that a multiplicity is needed to give a well-rounded view. Each of the schools (except, in fact, Göteborg) led the group of five cities on at least one of the criteria. With regard to the subfields in which Göteborg was strong or relatively weak, Fig. 7 shows that the two methods often disagreed: however they were united in revealing cardiovascular research, paediatrics, renal medicine and epidemiology as strengths; and haematology, genetics, cytology/cell biology and oncology as weaknesses. These conclusions were helpful to the panel when it was framing its recommendations.

## References

ANDERSON, J. (1989), The evaluation of research training. In: D.C. EVERED, S. HARNETT (Eds), *The Evaluation of Scientific Research*, 93. Wiley, Chichester.

ANDERSON, J., WILLIAMS, N., SEEMUNGAL, D., NARIN, F., OLIVASTRO, D. (1996), Human genetic technology: exploring the links between science and innovation, *Technology Analysis & Strategic Management*, 8: 135.

BIENENSTOCK, J., AF MALMBORG, C., HUTTUNEN, J., RODRIGUEZ-FARRÉ, E. (1996), *Evaluation of the Göteborg University Faculty of Medicine,* University of Göteborg, Sweden.

COLLINS, P.M.D. (1991), *Quantitative Assessment of Departmental Research*, SEPSU Policy Study no 5, The Royal Society, London.

DANIEL, H.-D., FISCH, R. (1990), Research performance evaluation in the German university sector, *Scientometrics*, 19: 349.

JESCHIN D., LEWISON G., ANDERSON J. (1995), A bibliometric database for tracking acknowledgements of research funding. In: *Proceedings of the 5th International Conference of the International Society for Scientometrics & Informetrics,* 235. Learned Information Inc., Medford NJ.

LEWISON, G. (1995), Evaluation of national biomedical research outputs through journal-based esteem measures, *Research Evaluation*, 5: 225-235.

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.

MARTIN, B.R. (1996), The use of multiple indicators in the assessment of basic research, *Scientometrics,* 36: 343.

NARIN, F., HAMILTON, K. S. (1996), Bibliometric performance measures, *Scientometrics*, 36: 293.

NARIN, F., PINSKI, G., GEE, H. H. (1976), Structure of the biomedical literature, *Journal of the American Society of Information Science,* 27: 25.

VAN RAAN, A. F. J. (1993), Advanced bibliometric methods to assess research performance and scientific development: basic principles and recent practical applications, *Research Evaluation*, 3: 151.