ORIGINAL INVESTIGATION

# Journal Impact Factors Do Not Equitably Reflect Academic Staff Performance in Different Medical Subspecialties

*Richard J. Epstein*


## ABSTRACT

**Background:** The simplest variables to quantify on an academic curriculum vitae are the impact factors (IFs) of journals in which articles have been published. As a result, these measures are increasingly used as part of academic staff assessment. The present study tests the hypotheses that IFs exhibit patterns that are consistent between journals of different specialties and that these IFs reflect the quality of staff academic performance.

**Methods:** The IFs of a sample of journals from each of four medical specialties—medicine, oncology, genetics, and public and occupational health—were downloaded from the *Science Citation Index* and compared. Overall and specialty-specific journal IF frequencies were analyzed with respect to distribution patterns, averages, and skew.

**Results:** Approximately 91% of journal IFs fell within the 0 to 5 range, with 97% being less than 10. The overall IF distribution featured a positive skew and a mean of 2.5. Separate analysis of the journal specialty subsets revealed significant differences in IF means (genetics 3.4 > oncology 3.1 > medicine 2.0 > public health 1.6; $p < .006$), all of which well exceeded the respective IF medians. Journals from the general medicine category exhibited both the lowest IF median (0.7) and the most positively skewed distribution.

**Conclusion:** The distribution of IFs exhibits degrees of skew, numeric average, and spread that differ significantly between journal specialty subsets. This suggests that factors other than random variations underlie much of the IF variation between specialty journals and reduces the plausibility of a reliable correlation between IFs and the quality of academic staff performance. It is concluded that a dominant emphasis on IFs in academic recruitment and promotion may select for long-term faculty characteristics other than academic quality alone.


**Key Words**: medical journals, academic medicine, bibliometrics

*From the Department of Medicine, University of Hong Kong, Queen Mary Hospital, Hong Kong.*

*Address correspondence to: Dr. Richard J. Epstein, Department of Medicine, University of Hong Kong, Professorial Block, Queen Mary Hospital, Pokfulam Road, Hong Kong; e-mail: repstein@hku.hk.*

For reasons that are primarily economic, global academic medicine now finds itself in a weaker position than in earlier decades.[1,2] One cause of this malaise is that the increasing dependence of academia on profit-oriented industrial collaborations[3] has weakened the public's previously powerful support for medical research.[4] A related problem is that the sink-or-swim nature of the modern academic career track has spawned the need for an efficient culling mechanism, as implied by widespread acceptance of the publish-or-perish ethos.[5]

Although most leading institutions have long recognized the counterproductive effects of basing academic promotion on the number of publications,[6] the pendulum has now swung to the other extreme: measures of quality have become the priority, reflecting the contemporary view that a performance-based "tipping point" must be attained to maximize the probability of tangible institutional returns, such as patents, royalties, spin-offs, joint ventures, and/or favorable media coverage.[7] The availability of a valid quality measure for individual research outputs would therefore prove welcome because this could relieve such institutions of having to agonize over complex issues, such as thematic research priorities, strategic investment, talent recognition schemes, and the development of sustainable academic career structures.

Since its creation in 1961, the US-based Institute for Scientific Information has derived an annual *Science Citation Index*, which, in turn, is used to generate a list of journal impact factors (IFs).[8] For example, a journal's IF for 2010 is calculable by counting the number of times that all articles from that journal published in 2008–2009 are cited by all journal articles in 2010 (citation number, C) and then dividing this frequency by the total number of articles published by the journal over the same 2-year period (article number, A). The IF thus represents an average annual citation rate for articles published by a specified journal; in general, the higher a journal IF, the higher is the assumption of peer review rigor and competition for acceptance,[9] leading, in turn, to higher assumptions of quality for academic work accepted for publication therein. It is this set of assumptions, then, that underlies the popular equation of IF with work quality and professional impact.

Many universities and similar biomedical institutions now rank and reward academic staff and departments mainly on the basis of published journal article IFs,[10] implying that the journal IFs attributed to a given staff member relate directly to the quality (Q) of his or her published work output.[11] In this way, IFs have become central to decisions about resource allocation and promotion in tertiary institutions, not least, it can be argued, because of their ready quantification. Despite this trend, many biases (Table 1) have been identified that cast doubt on the equation of IFs with work quality[12]; these shortcomings have generated many well-argued criticisms as to the overuse of IFs in the assessment of academic performance.[13–17] Nonetheless, because the interdependence of IFs and grant income influences recruitment decisions more as funding shrinks, faculties may expect to become even more enriched in "high-IF" staff characteristics irrespective of whether these equate with work quality alone.

Most quantifiable biologic traits—for example, IQ, memory, reading ability, job satisfaction, or body weight[18]—display a range of values approximating a normal (gaussian, bell curve) distribution. Hence, if the "quality" of a journal article reflects the research talents and weaknesses of its authors, it is reasonable to hypothesize that an accurate measure of article quality might also be normally distributed, with the mean and median values being similar. If the distribution is not normal, however, this implies that what is measured is less likely to reflect a biologic trait; an example would be the highly asymmetric distribution of wealth in the population, arguably illustrating that "it takes money to make money." Assuming that biomedical journal IFs bear a direct and consistent relationship to the quality (Q) of the average work published, then IFs should be normally distributed and should be capable of being represented in a manner that is internally consistent and reproducible between different fields of biomedicine. Given that the method for calculating IFs is simple arithmetic (IF = C/A), there is no a priori reason to reject the hypothesis that the IF value attributed to a given staff member's publication output (using whatever formula) reflects or approximates Q. Moreover, because it is reasonable to postulate that journals publish a spectrum of work ranging in quality from "very poor" to "very good," it is plausible that this spectrum resembles a normal distribution, that is, it is possible to identify and assign a midpoint such that similar numbers of "better" and "worse" journals have similarly deviated IFs on either side of this midpoint value. This finding would imply a smooth (ie, continuous) spectrum of quality differences between submitted outputs because the lower limit of the IF distribution is fixed at 0, whereas there is no upper limit; however, the distribution may be expected to feature a right-sided "tail" asymmetry of higher IF values. Here I analyze a variety of IF distributions and thus determine whether the hypothesized equation of quality with quantity can be confirmed or refuted.

**TABLE 1 Factors that May Influence the Impact Factor of a Journal Independently of the Quality of the Published Articles**

| *Nonquality Variables Affecting IF* | *Mechanism of Effect* | *Illustrative Example(s)* |
|---|---|---|
| Language of publication | Non-English publications are read by a smaller audience | The original papers of Einstein, Freud, etc., would have been assigned a low IF |
| Review articles | Citation reduces author effort in citing individual studies | *Bioessays*, *Adv Cancer Res*, *Annu Rev Biochem* |
| Methods articles | Cited as a technical shorthand, irrespective of the contribution's novelty or importance | The world's most highly quoted article, at 120,000 citations[25] |
| "Hot topics" | Quoted more often than less fashionable fields of work | SARS articles in 2003–2004 |
| Aggressive journal distribution or sale, including free electronic access | Increases readership, exposure, and hence citability | Supported by sponsorship, advertising, or professional body |
| Rapid publication of a journal | Increases citations as calculated, relative to less efficient journals[26] | Journals boasting "rapid review" process (eg, using a single reviewer or editor) |
| Large multiauthor publications | Proportionately greater use of self-referencing by the authors | Reports of cooperative clinical trials supported by industry |
| Subject matter | Scientific and technical articles contain more citations than clinical or humanities articles, generating greater "impact" for the field as a whole | Molecular biology or biochemistry journals compared with more lightly referenced journals |

IF = impact factor; SARS = severe acute respiratory syndrome.

## METHODS

All current biomedical journal IFs were electronically sourced via library subscription to the *Science Citation Index* 2003 database. The IFs of 400 biomedical journals were downloaded from the Institute for Scientific Information Web site via the University of Hong Kong intranet access. These journals comprised four subgroups: oncology and cancer (top 100 journals), medicine (top 100 journals), genetics and heredity (top 100 journals), and public and occupational health (all 89 journals). The range of current journal IFs extended from C/A values of 0 (ie, not cited in the last 2 years) to 33 (ie, average article published in the journal in question over the last 2 years is cited 33 times in the current year). Analyses were undertaken using the entire dataset and the subject-specific journal subgroups.

Because the IFs, as calculated, represent continuous rather than discrete variables, no modal distribution could be determined; for this purpose, IF histogram ranges were used. To improve the clarity of graphic analysis, the 3.25% of journal IFs exceeding 10 (ie, rare high outliers) were excluded from graphic analyses (but see Table 2). The *Microsoft Excel* spreadsheet program was used to determine distribution means, standard deviations, and other statistical variables.

In view of the non-normal distributions involved (see below), both the nonparametric Kruskal-Wallis test and the median test were used to quantify the significance of differences in IF distributions and their medians, respectively, between journals from different specialty subsets, in preference to *t*-testing.

## RESULTS

The majority of indexed journals (96.8%) are characterized by IFs within the range of 0 to 10 and 91.3% by IFs within the 0 to 5 range, corresponding to the first 10 labeled histograms on the graph, each of which represents half of an

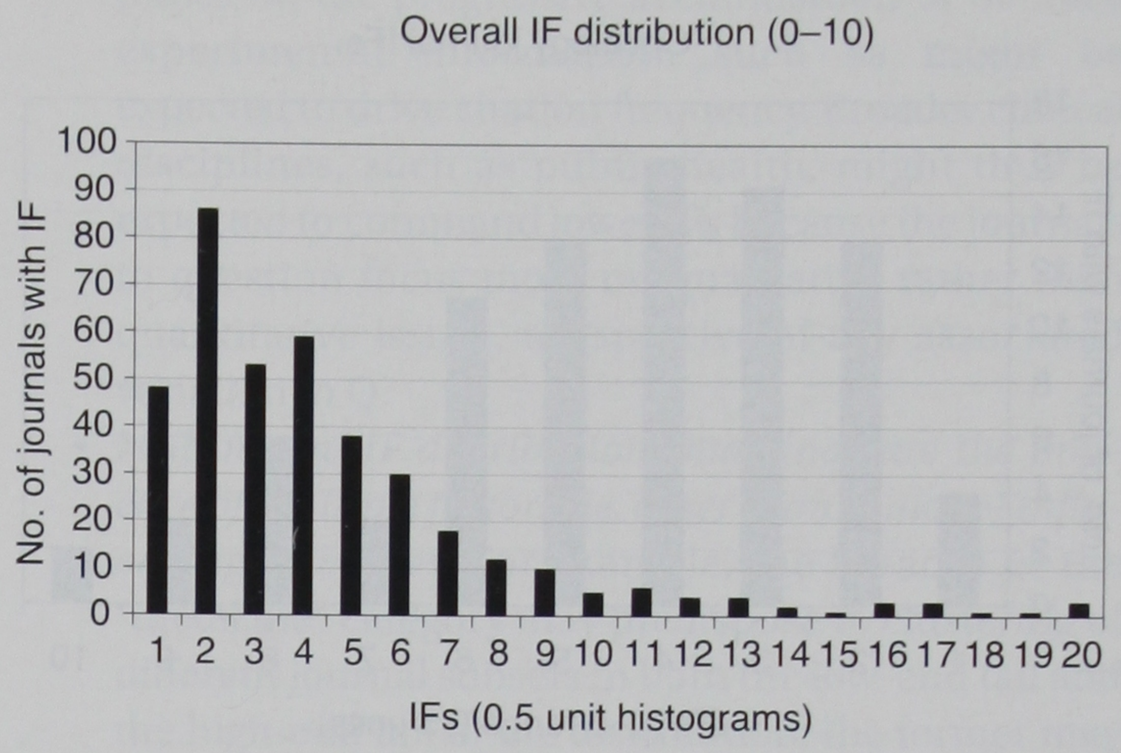


FIGURE 1 The overall impact factor (IF) distribution of the journals surveyed. Each histogram corresponds to a 0.5 IF unit range.

IF unit spread (Figure 1). The distribution pattern of IFs is not normal but skewed, and the distribution is discontinuous: the primary modal range is represented by the second histogram (IF range 0.5–0.99), whereas the numeric mean of this $0 < IF < 10$ distribution lies within the secondary modal range (histogram 4).

Subgroup analysis presented in Figure 2 reveals that the overall distribution represents the sum of disparate patterns. Compare, for example, the oncology and medicine journal IF spectra: for the former, the range-restricted IF histogram frequency density approximates a normal distribution, but this is not the case for medicine journals. The discontinuous (multimodal) IF distribution of the genetics journals is difficult to interpret given the relatively small sample size; it does raise the possibility, however, that factors other than random variations in quality may influence IF allocations.

Variations in distribution between the different specialist journals were quantified as shown in Table 2 (the amended mean and standard deviation, corrected for the entire raw data set, including IFs greater than 10, are itali-

**TABLE 2 Variations of Impact Factor Pattern and Size as a Function of Journal Area of Specialization (Oncology, Medicine, Genetics, or Public/Occupational Health)**

| *Journal Specialty* | *IF (0–10) Median (IF 0–33)* | *IF (0–10) Modal Range* | *IF (0–10) Mean (IF 0–33)* | *IF (0–10) SD (IF 0–33)* | *IF (0–10) Skew (IF 0–33)* |
|---|---|---|---|---|---|
| Oncology (*n* = 100) | 2.00 (*2.10*) | 1.5–1.99 | 2.55 (*3.08*) | 2.02 (*3.96*) | 1.81 (*4.98*) |
| Medicine (*n* = 100) | 0.70 (*0.75*) | 0.0–0.49 | 1.27 (*1.97*) | 1.49 (*4.06*) | 2.69 (*5.10*) |
| Genetics (*n* = 100) | 2.00 (*2.20*) | 1.5–1.99 | 2.62 (*3.41*) | 2.15 (*4.23*) | 1.67 (*3.32*) |
| Public health (*n* = 89) | 1.30 (*1.30*) | 0.5–0.99 | 1.56 (*1.56*) | 1.21 (*1.21*) | 2.15 (*2.15*) |
| All | 1.55 (*1.60*) | 0.5–0.99 | 1.99 (*2.51*) | 1.85 (*3.65*) | 2.04 (*4.76*) |

Values for impact factors (IFs) less than 10 are presented in roman typeface, whereas values in parentheses are used to designate calculations based on the entire dataset.

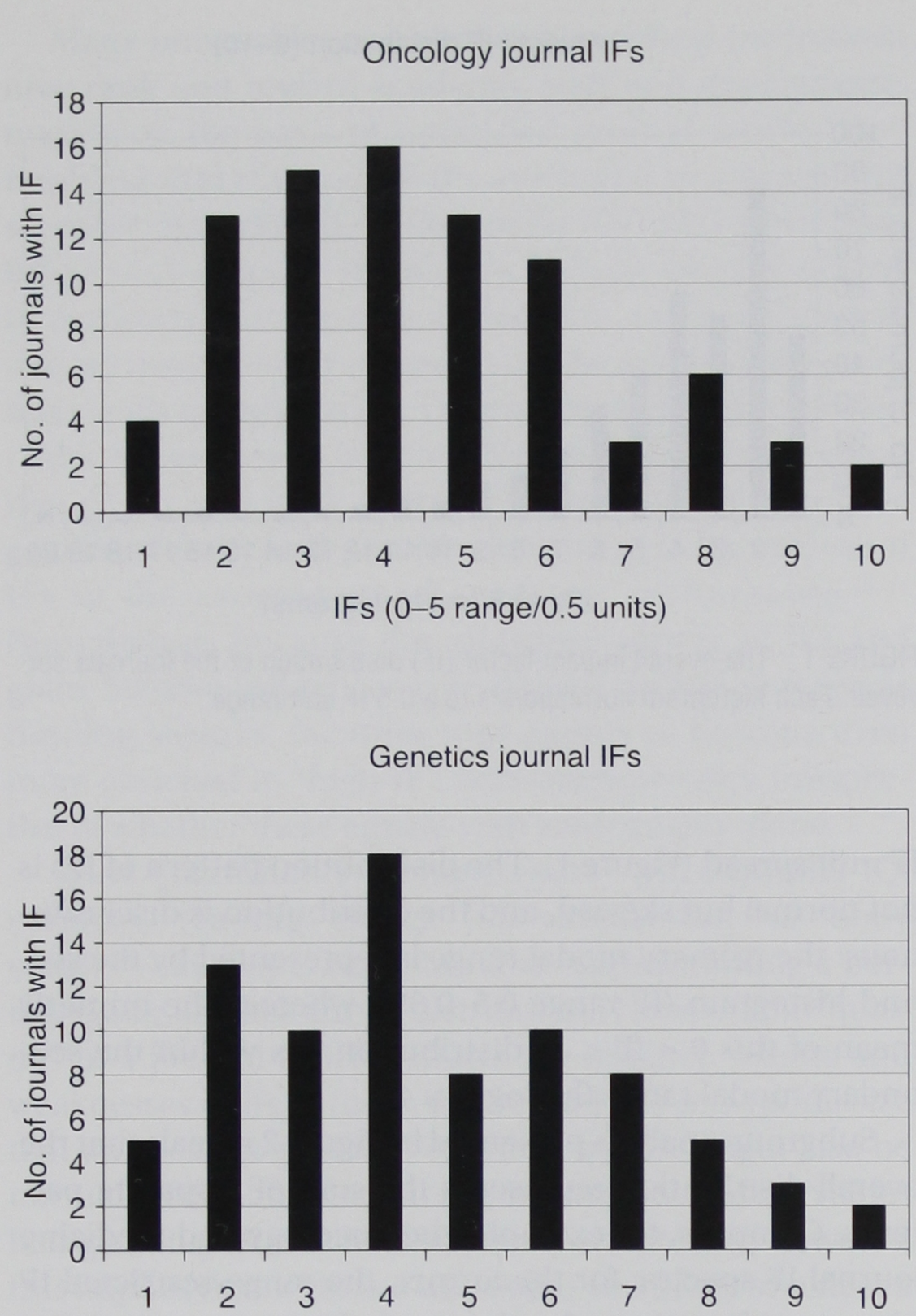


FIGURE 2 Side-by-side comparison of distribution patterns between different journal subspecialties. IF = impact factor.

cized in parentheses for comparison). Oncology and medicine IFs differ in terms of their midpoints—the oncology IF mean being double that of medicine—and of their distribution shape. With respect to the latter, the fact that the medicine IF standard deviation exceeds the mean is explained by the pronounced positive (right) skew of the distribution. Nonparametric Kruskal-Wallis testing of mean rank confirms a significant difference between the subspecialty IFs examined ($p < .006$; $\chi_2$ 12.65).

Evidence of a right skew to all of the sample subsets is indicated by the medians being less than the mean,[19] and application of the median test again confirms a highly significant difference between the journal subspecialty IFs in this respect ($p < .009$; $\chi_2$ 11.73). As confirmed by the skew constant in column 6, the extent to which the IF distributions are asymmetrically distributed varies between specialties such that medicine > public health > oncology > genetics. The sharp growth of the skew constants when the range of measurements is increased to include the outliers (italicized in parentheses below) reflects the distribution becoming more "left triangular," that is, positively skewed.

As expected, inclusion of outliers (IFs > 10) greatly increases the standard deviation, most notably in medicine, where this measure of average spread more than doubles (there is no change in public health because it has no journals with an IF exceeding 10). In contrast, the median changes little in response to this extension of the range by over 200%, which is characteristic of distributions distorted by a small number of outliers. Based on this retrospective analysis, then, differences in IF distribution as a function of journal subspecialty follow certain nonrandom rules: (1) more prevalent low-IF journals: medicine > public health > oncology > genetics; (2) more prevalent high-IF journals: genetics > oncology > medicine > public health; and (3) higher mean or median IF journals: genetics > oncology > public health > medicine.

## DISCUSSION

The central focus of this analysis is the relationship, if any, between the specific field of an academic medical journal and its calculated IF. As shown above, such relationships do exist but are neither predictable nor consistent: patterns of journal IF size, spread, and skew differ substantially between the specialty fields examined but without any explanatory correlation discernible (eg, public health

may be argued to be the most "important" field in terms of its potential medical impact, but it has the lowest IF). Another way to make this point is to consider journals with a specific 2003 IF—say, in the range 4.2 to 4.4—from each of the fields surveyed: the intraspecialty rank order of such journals is 22, 20, 10, and 4 for genetics, oncology, medicine, and public health, respectively (data not shown).

A more general issue addressed by this study is whether the quantitative relationship between Q and IF is consistent with the hypothesis that raw IF data may be used in practice as surrogates for more labor-intensive Q estimations. However, several aspects of the present data set suggest that the relationship between Q and IF is neither simple nor direct; to appreciate this, let us first consider which pattern of results would support the test hypothesis. First, given that any quantification of Q will have a median value—with 50% of all of the sample's Q values being above and 50% below this value—the simplest relationship for a putative covariable such as IF will be that of a normal distribution, with the mean and median values closely coinciding. Second, because the fundamental variables governing the quantification of Q might reasonably be assumed to be similar across different biomedical subspecialties, the pattern of variation between Q and any covariable (such as IF) should also be similar in different subsets. Third, in the absence of prior evidence for systematic differences in the quality of academic outputs from different specialties, only very small quantitative differences in any putative Q covariables (such as IF medians, means, and standard deviations) would be expected between these sample subsets. As noted above, however, none of these criteria are fulfilled by the results of this analysis. What factors might account for this apparent confounding? Reverting to the raw data, the following possibilities are suggested:

- *Different classes of biomedical journal have dissimilar audience sizes.* A striking feature of this analysis relates to the highly asymmetric, right-skewed, unimodal IF distribution of medicine, which happens to be a somewhat generalist classification rather than a discrete, well-focused research area, such as oncology or genetics. On reviewing the journal titles in this category, many are revealed to be the official organs of professional societies rather than research-oriented journals of record. The circulation of such journals may thus be predictably (even intentionally) restricted, thereby reducing the IF to an extent that may be disproportionate to any associated Q reduction.
- *Differences in research detail and complexity between fields may manifest as a decline in IF independent of Q.* Of the three "specialty" journal classes represented, median and mean IFs decline in the order genetics > oncology > public health. Arguably, this order also reflects the reliance of each of these disciplines on the progressive accumulation of detailed experimental information, such as might be expected to drive citation frequency. Broader clinical disciplines, such as public health, might thus be expected to command lower IFs because the journals in question focus more on qualitative rather than quantitative issues, irrespective of any associated variation in Q.
- *Multimodal IF distributions may indicate the presence of multiple factors (ie, other than Q alone) influencing IF values.* For example, the breadth of the "medicine" category may predispose to clustering of different journal subsets in both the low-end tail and the high-end tail of the distribution; the former may include journals that are too general in orientation for research citation, whereas the latter may include the most groundbreaking studies relevant to the broadest biomedical audiences. Alternatively, rapidly advancing ("hot") fields of research, such as genetics, may be characterized by high or rising IFs but may also spawn large numbers of new market entrants; these newer journals will initially tend to cluster at the lower end of the IF distribution and dilute the median and mean IF rankings of the field, independent of present or future Q.

These conclusions strengthen and extend the concerns raised by many others concerning the hazards of overreliance on impact factors as academic quality surrogates.[8,13,16,17] In an analysis of 204 published articles, Callaham and colleagues found that traditional "quality" measures of study methodology and design were poorly predictive of both citation frequency (ie, impact) and the IF of the publishing journal; in fact, the only variable to prove highly predictive of article citation frequency was the IF of the publishing journal, suggesting that chance or nonquality variables relating to manuscript acceptance (eg, newsworthiness as perceived by the journal or the persistence or "gamesmanship" of an author[20]) may determine the ultimate impact of a submission.[21] In addition, Rostami-Hodjegan and Tucker demonstrated highly skewed citation frequency distributions of individual articles within journals of a given IF, casting further doubt on the interpretability of IFs as a measure of individual research output.[22] Healy and Cattell reported that industry-linked studies on a given subject tend to have a sixfold higher IF than non–industry-linked studies, raising potential concerns about the long-term objectivity and interpretability of the present system.[23]

Inevitably, there are many limitations of a small cross-sectional analysis such as the present study. For example, it is well recognized that journal IFs change over time; if public health is a newer academic field than genetics or oncology, for the sake of argument, it may be expected that the 2003 IF will be temporarily lower. Moreover, this study does not address the heterogeneity of articles within a

given journal—an important limitation of the IF "system," as pointed out by other authors[24]—although the number and choice of the specialty fields selected are arbitrary, limited as they are to the author's interests. Perhaps the most contentious assumption of this study, however, is the notion that journal "quality" (defined as Q) should produce a normal distribution of IFs, assuming, in turn, that IFs are indeed surrogates for quality measurement. This proposal reflects the view that no irrefutable (objective) measure of (subjective) quality exists; on the other hand, most random variations of a human trait will tend to adopt a normal distribution (see the introduction). Given that, in this analysis, the variances of IFs are not normally distributed either within or across specialties, it is concluded that IFs are unlikely to correlate closely with the "quality" of the authors' academic talents and that other variables are likely to exert a major influence on the quantitation of IF (see Table 1).

Not all medical schools use IFs in the same way to assess or grade academic staff, and many institutions are keenly aware of the limitations of this practice. Nonetheless, tacit acceptance of IFs as a marker of academic performance undoubtedly remains widespread and seems likely to increase the darwinian selection pressure on researchers to optimize their IF-seeking skills (see Table 1) at the inevitable expense of competing priorities, such as teaching and mentoring, administrative leadership, collegiality, advocacy, and integrity. Hard work (such as might oblige a recruitment committee to read and assess a candidate's publications), strategic goal setting, innovative appraisal systems, and more diversified streams of government support may be needed to slow the attrition of these traditional academic values.

## ACKNOWLEDGMENT

I thank Dr. Yongzhong Zhao for statistical assistance.

## REFERENCES

1. Tugwell P. Campaign to revitalise academic medicine kicks off. BMJ 2004;328:597.
2. Goldbeck-Wood S. Reviving academic medicine in Britain. BMJ 2000;320:591–2.
3. Korn D. Industry, academia, investigator: managing the relationships. Acad Med 2002;77:1089–95.
4. Clark J. Polishing the tarnished image of academic medicine. BM J 2004;328:604.
5. Pearson H. It's a scoop! Nature 2003;426:222-3.
6. Angell M. Publish or perish: a proposal. Ann Intern Med 1986;104:261–2.
7. Gladwell M. The tipping point. New York: Little, Brown & Co.; 2003.
8. Opthof T. Sense and nonsense about the impact factor. Cardiovasc Res 1997;33:1–7.
9. Lee KP, Schotland M, Bacchetti P, Bero LA. Association of journal quality indicators with methodological quality of clinical research articles. JAMA 2002;287:2805-8.
10. Hecht F, Hecht BK, Sandberg AA. The journal "impact factor": a misnamed, misleading, misused measure. Cancer Genet Cytogenet 1998;104:77–81.
11. Saha S, Saint S, Christakis DA. Impact factor: a valid measure of journal quality? J Med Libr Assoc 2003;91:42-6.
12. Ojasoo T, Maisonneuve H, Matillon Y. The impact factor of medical journals, a bibliometric indicator to be handled with care. Presse Med 2002;31:775–81.
13. Seglen PO. Why the impact factor of journals should not be used for evaluating research. BMJ 1997;314:497.
14. Lankhorst GJ, Franchignoni F. The 'impact factor'—an explanation and its application to rehabilitation journals. Clin Rehabil 2001;15:115-8.
15. Van Diest PJ, Holzel H, Burnett D, Crocker J. Impactitis: new cures for an old disease. J Clin Pathol 2001;54:817-9.
16. Kurmis AP. Understanding the limitations of the journal impact factor. J Bone Joint Surg Am 2003;85:2449–54.
17. Jones AW. Impact factors of forensic science and toxicology journals: what do the numbers really mean? Forens Sci Int 2003;133:1–8.
18. Archer ZA, Rayner DV, Rozman J, et al. Normal distribution of body weight in male rats fed a high-energy diet. Obes Res 2003;11:1376–83.
19. Swift L. Quantitative methods. Basingstoke, Hampshire, UK: Palgrave Macmillan; 2001.
20. Sindermann CJ. Winning the games scientists play. Cambridge (MA): Perseus Publishing; 2001.
21. Callaham M, Wears RL, Weber E. Journal prestige, publication bias, and other characteristics associated with citation of published studies in peer-reviewd journals. JAMA 2002; 287:2847–50.
22. Rostami-Hodjegan A, Tucker GT. Journal impact factors: a ''bioequivalence" issue? Br J Clin Pharmacol 2001;51:111-7.
23. Healy D, Cattell D. Interface between authorship, industry and science in the domain of therapeutics. Br J Psychiatry 2003;183:22-7.
24. Weale AR, Bailey M, Lear PA. The level of noncitation of articles within a journal as a measure of quality: a comparison to the impact factor. BMC Med Res Methodol 2004;4:14–22.
25. Feinberg AP, Vogelstein B. A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. Anal Biochem 1983;132:6–13.
26. Jones AW. JAT's impact factor—room for improvement? J Anal Toxicol 2002;26:2–5.