subject category "biology", it is thus placed at the number 1 position of 64 in its first year of reporting an IF. FASEB journal at position 2 has an IF of 6.8, but has been in circulation since 1987. Similarly, in the other SCI subject category ("biochemistry and molecular biology")in which PLOS Biology is listed, it ranks at position 8 out of 261. Monitoring the development of such journals' IF will inform the determination of the online-availability bias in the future. This effect will increase in the future with the availability of new search engines with deep penetration such as Google Scholar [25,26], allowing researchers to find relevant articles in an instant, and then choose those with immediately and freely available content over those with barriers, economic and otherwise.

Accuracy of data capture by ISI

Investigations by *Nature* suggested a significant undercount of "citable" items in *Nature Genetics* in 1996 and an erroneous inclusion of "citable" items other than those defined by ISI itself for *Nature* in 2000 [4]. A more recent issue is undercounted citations to articles authored by consortia, rather than by a list of individual authors [27]. The article reporting the draft human genome sequence from the International Human Genome Sequence Consortium [28] is considered as a landmark paper published in *Nature* in 2001, but was surprisingly absent from the list of "hot papers" in biology, which are published regularly by ISI Science Watch [29]. The examination of the ISI's data showed that the ISI only considered citations to the full list of authors, led by Eric Lander of the Whitehead Institute for Biomedical Research at Cambridge, Massachusetts, and hence led to the grossly undercounted representation. The same applied to other prominent papers authored by consortia.

The accuracy of how citations are collected at ISI significantly influences the final published IF statistics. At ISI, data capture of journal papers is completed by optical character recognition software, important fields are highlighted manually, and the final tagging of every individual article is computerized. Algorithms have been designed to count valid citations. The simple involvement of two highlighted fields, such as the journal title and the year, makes the citation counting for the IF calculation easier. However, this raises systematic bias as citations cannot be matched to individual articles in real time. At this point an incorrect citation leaves its mark. For the field of environmental and occupational medicine, a recent study reported a prevalence of 3.35% incorrect citations; the respective articles receive an incorrect cite count, thus potentially reducing their journals' IF [30].

IF is calculated for a whole journal whereas citations are to individual articles

The IF would reflect a journal's interest to the research community if citations were indeed distributed equally over all articles in the journal. However, this is not the case. Only a small percentage of articles are highly cited. Based on the analysis of three biochemical journals, Seglen [13] found that the most cited 15% of articles account for 50% of the citations and the most cited 50% of articles account for almost all citations (90%). These numbers were confirmed by a later study based on two cardiovascular journals [31]. The most recent study on articles published in *Nature* showed a similar high skew of citations: 89% of 2004's citations were generated by just 25% of *Nature*'s papers [32]. Apparently, researchers cannot solely depend on the IF to judge the quality of the journal.

Highly cited articles are found mostly in a small subset of journals, regardless of how parameters of the algorithm (e.g. average time-frame) are changed. In Garfield's view, these two combined effects strengthen the ISI's position as a means to point authors and readers to journals with true scientific impact [8]. The argument is that this effect justifies the fact that JCI is not all-inclusive, but rather selective. According to Garfield, JCI could still be considered comprehensive if it covered only the 500 most cited journals.

Invalid articles may pose a considerable bias on the journal IF. Retracted articles may continue to be cited by others as valid work. Pfeifer and Snodgrass [33] identified 82 completely retracted articles, analyzed their subsequent use in the scientific literature, and found that these retractions were still cited hundreds of times to support scientific concepts. Kochan and Budd [34] showed that retracted papers by John Darsee based on fabricated data were still positively cited in the cardiology literature although years had passed since retraction. Budd et al. [35] obtained all retractions from MEDLINE between 1966 and August 1997 and found that many papers still cited retracted papers as valid research long after the retraction notice.

Interesting papers, based on fraudulent data, may attract the scientific community's attention and be cited frequently, thus distorting the true impact of the journal that featured the sensational article. In a notable 2002 case of scientific fraud, Jan Hendrik Schön, a former researcher at Bell Laboratory, published "remarkable" findings on superconductivity, molecular electronics, and molecular crystals in several scientific journals, including *Science*, *Nature* and *Applied Physics Letters*. He was later found out to have falsified or fabricated data in 16 of 24 alleged cases of misconduct [36]. The data of 25 publications were implicated in the perpetuation of dubious claims. The