SUPPLEMENTARY INFORMATION

Mathematical model to calculate the number of articles that should have been retracted

We analysed the PubMed database, looking for retracted articles published between 1950 and 2004 in 4,348 journals with known impact factors (*IF*s) over a number of years. Here we compute each journal's impact factor as an average of its *IF* values reported for the years 1999 to 2004 by ISI Thomson, Inc. Note that for some journals, ISI provides different *IF* values in the bar charts and text on their website; we have used the values described in the text.

Let IF_i be the impact factor of the i^{th} journal, and IF_{max} be the highest impact factor that we observe in our dataset (50.551). We define a normalized *IF* for the i^{th} journal, r_i, as $r_i = \frac{IF_i}{IF_{\max}}$. Let a_i be the total number of articles published in the i^{th} journal and ψ_i be the number of retracted articles in the same journal. According to our model outlined in Figure 1C (inset), the probability of observing retraction of ψ_i out of a_i articles published in the i^{th} journal, computed jointly for all N journals in our dataset (i=1,…,n), is

$$p\left(\{a_i,\psi_i,IF_i\}\,|\,\Theta\right)=\binom{\sum_i a_i}{a_1\ldots a_n}\prod_i p(IF_i)^{a_i}\prod_i\left\{\binom{a_i}{\psi_i}\left[\left(1-\theta r_i^{\alpha}\right)\cdot\tau r_i^{\beta}\right]^{\psi_i}\left[1-\left(1-\theta r_i^{\alpha}\right)\cdot\tau r_i^{\beta}\right]^{(a_i-\psi_i)}\right\}. \quad (1)$$

In this expression $p(IF_i)$ is the probability of sampling an article that is published in the i^{th} journal (with impact factor IF_i). Note that the multinomial probability of sampling the whole

observed article set, $\binom{\sum_{i=1}^{n} a_i}{a_1 \ldots a_n} \prod_{i=1}^{n} p(IF_i)^{a_i}$, and the binomial coefficient, $\binom{a_i}{\psi_i}$, do not depend on the values of our model parameters and, therefore, they can be omitted in the maximum likelihood and MCMC computations.

A set of 5 journals from the ISI dataset have IF assigned to 0. Most certainly the articles published in these journals are cited somewhere, but these citations fall outside of the set of journals reviewed by the ISI. We attempt to account for this incompleteness of the ISI data in our calculation of impact factors in the following way. For the set of 5 journals with ISI-assigned IF of 0, we postulate a pseudo-IF of 0.0009, one tenth of the smallest IF that we observe in our dataset.

In our model, parameters θ and τ can vary between 0 and 1, while α and β can take any real value. We require that the joint probability of θ, α, τ, and β be 0 whenever θr_i^{α} or τr_i^{β} are smaller than 0 or larger than 1, because we define these quantities as probabilities.

To estimate the posterior distribution of parameter values (given an uninformative prior distribution over parameter values), we use Markov chain Monte Carlo (MCMC; Gilks *et al*, 1996). We repeat our parameter estimation while fixing τ to several values between 0 and 1, using MCMC with 10 million iterations. The results of this parameter estimation are given in Figure 1B.

The expected number of retractable articles for particular values of α and θ can be computed as:

$$R_i = a_i(1-\theta r_i^{\alpha}), \quad (2)$$

$$R = \sum_{i=1}^{N} R_i, \quad (3)$$

where R is the total (unobserved) number of retractable articles, R_i is the i^{th} journal's share of this number, and n is the total number of journals. The parameters a_i and r_i denote the number of articles and the normalized *IF* of the i^{th} journal, respectively. We use the joint posterior distribution of α, β, and θ for τ=0.1 *and* τ=1 values to compute the posterior distributions of R and R_i shown in Figure 1C.

ACKNOWLEDGEMENTS

The authors are grateful to Ms. Rebecca Baldridge and Dr. Hülya Yazıcı for comments on the earlier version of the manuscript. This work was supported by the US National Institutes of Health.

REFERENCE

Gilks WR, Richardson S, Spiegelhalter DJ (eds; 1996) *Markov chain Monte Carlo in practice*. New York, NY, USA: Chapman & Hall