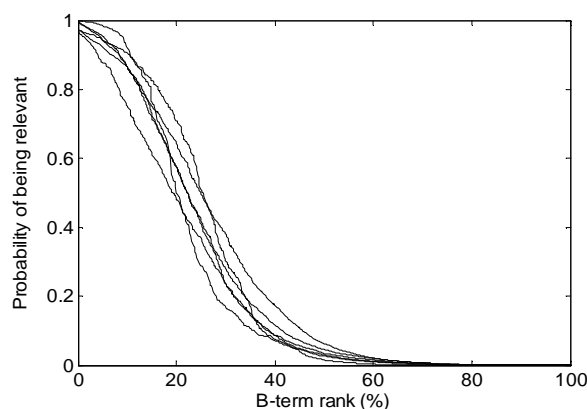


# A quantitative model for linking two disparate sets of articles in MEDLINE

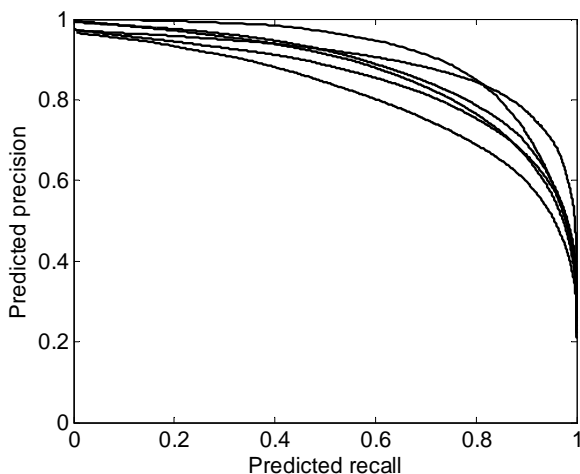
Vetle I. Torvik and Neil R. Smalheiser\*

Department of Psychiatry and Psychiatric Institute (MC912), University of Illinois-Chicago, Chicago, IL 60612 USA

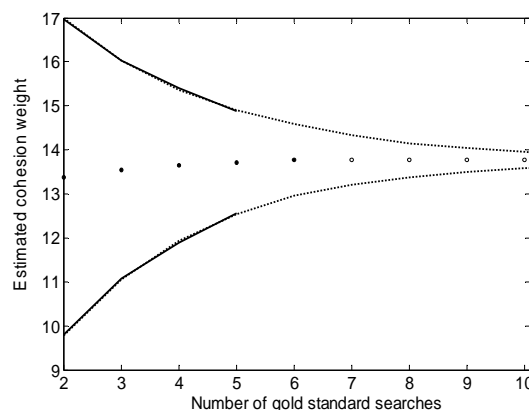
## SUPPLEMENTARY INFORMATION



**Fig. S1.** Relationship between the rank of a B-term and its probability of being relevant. Each curve corresponds to a gold standard two-node search. B-terms were ranked by score; the rank number of each B-term was then divided by the total number of B-terms, and multiplied by 100 to give its percentile rank.



**Fig. S2.** Precision/recall curves for the six gold standard two-node searches. Precision at a given B-term score is defined as the proportion of B-terms having that value or greater that are predicted to be relevant. Recall is defined as the predicted number of relevant B-terms having that value or greater, divided by the total predicted number of relevant B-terms.



**Fig. S3.** The uncertainty in estimating parameter weights decreases as the number of gold standard two-node searches in the model increases. The estimated weight of the cohesion feature  $w_4$  (Supplementary Table S1) is plotted for subsets of 2, 3, 4 and 5 gold standard two-node searches for logistic regression and are extrapolated up to 10 gold standards (the other feature weights behaved similarly). The curves show the standard deviation about the mean.

**Table S1.** Eight features used to characterize each B-term.

<p><b>1. Does the B-term occur in more than one paper within literatures A and C?</b></p> <p>Previous informal analyses suggested that removing B-terms that occur only once in A or C removed about 75% of the total number of B-terms but left most useful B-terms untouched. That suggested that a minimal frequency of occurrence might be one predictor of relevance. Terms which occur in only one title in A or C are penalized, but only if the corresponding literature has over 1,000 articles.</p> $x_1 = 1 \text{ if } (n_{AB} > 1 \text{ OR } n_A < 1000) \text{ AND } (n_{BC} > 1 \text{ OR } n_C < 1,000); 0 \text{ otherwise.}$
<p><b>2. Do the AB and BC sub-literatures share any MeSH terms?</b></p> <p>B-terms, in which the corresponding AB and BC papers do not share any Medical Subject Headings, are less likely to be relevant than those in which MeSH terms are shared (Swanson et al., 2006). Note that this feature is not solely determined by the nature of the B-term itself, but depends upon the specific query literatures A and C. Terms for which AB and BC do not share any MeSH are penalized.</p> $x_2 = 1 \text{ if there is at least one MeSH term in AB that is also in BC, [a MeSH term stoplist is applied of the most common and general MeSH terms]; 0.5 if none of the articles in AB (or BC) have assigned MeSH terms; 0 otherwise.}$
<p><b>3. Does the B-term map to at least one UMLS semantic category?</b></p> <p>In most two-node searches, the field testers were looking for B-terms that fell into specific, well recognized semantic categories, e.g., proteins or surgical procedures (Table 1). Terms that are not mapped to any UMLS semantic category (McCray et al., 2001) [using MMTx (Aronson, 2001)] are penalized. Terms that fail to map using MMTx but mapped to a list of neuroanatomical terms (Bowden and Dubach, 2003), or a list of gene and protein names (Tanabe and Wilbur, 2004), were also considered as mapping to semantic categories.</p> $x_3 = 1 \text{ if B mapped to a semantic category; 0 otherwise.}$
<p><b>4. Does the B-term have a high literature cohesion score?</b></p> <p>Literature cohesion measures how many different MeSH terms are used in a given literature compared to the overall size of the literature (Swanson et al., 2006). The literature cohesion score of a B-term is the cohesion score of the set of articles within MEDLINE that have the B-term in the title. The more topically specific and tightly focused the literature corresponding to a term is, the higher its score:</p> $x_4 = \min\{0.3, \text{coh}(B)\}$ <p>Here, <math>\text{coh}(B)</math> is the total count of the <math>k</math> top-most frequent MeSH terms (excluding the greatest count) divided by the total count of all MeSH terms within the set of papers in MEDLINE that have <math>B</math> in the title. The parameter <math>k</math> is a function of the size of the literature <math>u</math>: <math>k = \text{integer}(1.7 * \ln(u) + 0.5)</math>, which translates into 4, 8, 12, and 16 for <math>u = 10, 100, 1000</math>, and 10,000 papers, respectively (Swanson et al., 2006). The maximal value of <math>x_4</math> was capped at 0.3 to avoid nonlinearities observed at very high values.</p> <p>The 20 top-most frequent MeSH terms in MEDLINE were removed prior to computing the cohesion values.</p>
<p><b>5. Is the B-term moderately frequent within MEDLINE as a whole?</b></p> <p>Preliminary analyses suggested that relevant B-terms tended not to be extremely common nor extremely rare in general usage. Very frequent and very infrequent terms are penalized by 1 unit if <math>n_B = 100</math> or 10,000, by 2 units if <math>n_B = 10</math> or 100,000, etc.</p> $x_5 =  3 - \log_{10}(n_B) .$
<p><b>6. Did the B-term first appear recently within MEDLINE as a whole?</b></p> <p>A priori, it was not clear whether recency would be a factor in determining relevance. The more recently the term first appeared in MEDLINE, the higher its score:</p> $x_6 = \text{greater of } 1950 \text{ OR first year occurred in a MEDLINE title or abstract (assigned 2005 if not in the term db).}$

**7. Is the B-term highly characteristic within literature A or C?**

A B-term is said to be characteristic within a literature if it occurs significantly more frequently within titles in literature A or C than expected by chance, given the size of the literature and its overall frequency within MEDLINE as a whole. The more characteristic the term is within A or C, the higher its score:

$$x_7 = \min\{-\log_{10}(\text{p-value}), 8\},$$

where  $\text{p-value} = \sum (e^{-\lambda} \lambda^i / i!, i = n_{AB} + n_{BC}, \dots, \infty)$  is the Poisson approximation,  $\lambda = (n_A + n_C)n_B/N$ , of the binomial probability of observing  $(n_{AB} + n_{BC})$  or more successes out of  $(n_A + n_C)$  trials given a success probability of  $n_B/N$  for each trial. ( $N$  = total number of articles in MEDLINE.)

**8. Do the words within the B-term all occur on the customized 1400 word stoplist?**

A stoplist speeds up processing but brings the risk of removing relevant phrases (e.g., “Down syndrome” would be removed since both “down” and “syndrome” are stoplisted). Thus, we assessed whether use of a stoplist (longer than the short PubMed stoplist) would improve the overall performance of the system. Terms that are made up entirely of words on a customized stoplist (a list of the approximately 1400 most frequent words in MEDLINE; list can be viewed on the Arrowsmith website ([http://arrowsmith.psych.uic.edu/arrowsmith\\_uic/data/stopwords\\_1000](http://arrowsmith.psych.uic.edu/arrowsmith_uic/data/stopwords_1000))) are penalized.

$$x_8 = 0 \text{ if B is made up of entirely of stopwords; } 1 \text{ otherwise.}$$

**Table S2.** Parameter weights estimated by the logistic regression model and the pooled gold standard dataset. All the predictive features are statistically significant (p-value = 0.00003 or better).

Parameter	Estimate	Standard error (SE)	t-statistic (estimate/SE)
w <sub>1</sub>	0.732	0.16	4.71
w <sub>2</sub>	0.988	0.25	4.01
w <sub>3</sub>	1.317	0.26	5.10
w <sub>4</sub>	13.77	1.25	11.04
w <sub>5</sub>	0.586	0.11	5.11
w <sub>6</sub>	0.0396	0.0055	7.21
w <sub>7</sub>	0.189	0.025	7.52

**Table S3.** Two-node searches expected to have relatively little meaningful implicit information. They were adjusted to match roughly the sizes of the gold standard searches.

A-literature query	C-literature query	B-terms	Predicted relevant (%)
"weightlessness simulation"[mh] n = 1653	"aztreonam"[mh] n = 1104	655	8
"circadian rhythm"[mh] AND circadian rhythm*[ti] n = 4964	"sea urchins"[mh] OR sea urchins[tw] n = 6300	2917	9
"mesothelioma"[mh] AND mesothelioma*[tiab] n = 6179	"authoritarianism"[mh] OR authoritarianism[tw] n = 1915	1086	10
"smoking cessation"[mh] AND smoking cessation[ti] n = 2745	"grasshoppers"[mh] AND grasshopper[tiab] n = 580	534	12
"ampicillin resistance"[mh] n = 878	"euthanasia, active, voluntary"[mh] AND euthanasia[ti] n = 671	241	8
"lasers"[mh] AND lasers[tiab] n = 2420	"fluconazole"[mh] AND fluconazole[tiab] n = 3165	1434	16

**Table S4.** Two-node searches consisting of pairs of closely related topics.

A-literature query	C-literature query	B-terms	Predicted relevant (%)
"mesothelioma/etiology"[mh] n = 1144	"mesothelioma/physiology"[mh] n = 3132	1063	38
("migraine disorders"[mh] AND migraine[ti]) AND clinical trial[ptyp] n = 1248	("headache"[mh] AND headache[ti]) AND clinical trial[ptyp] n = 416	844	38
"parkinson disease/etiology"[mh] AND (parkinson[ti] OR parkinson's[ti]) n = 4147	"alzheimer disease/etiology"[mh] AND (alzheimer[ti] OR alzheimer's[ti]) n = 7739	5028	33
"rna processing, post-transcriptional"[mh] AND mammalian[All Fields] n = 1867	"rna interference"[mh] AND mammalian[All Fields] n = 620	1436	33
"subjective contour"[All Fields] OR "illusory contour"[All Fields] OR "subjective contours"[All Fields] OR "illusory contours"[All Fields] n = 332	(V1[All Fields] AND cortex[All Fields]) OR "striate cortex"[ti] OR "primary visual cortex"[ti] OR "area 17"[All Fields] n = 3805	711	27