

AnonMed: Sentence-level search engine for the MEDLINE database of biomedical articles

Mir S Siadaty*, Jianfen Shu, William A Knaus

Department of Public Health Sciences, University of Virginia School of Medicine, Box 800717, Charlottesville, Virginia, 22908, USA

*Corresponding author

Email addresses:

MSS: MirSiadaty@virginia.edu

JS: jshu@virginia.edu

WAK: wak4b@virginia.edu

Abstract

Background

Nowadays a substantial portion of the large amounts of data produced in different organizations is recorded in digital form. This enables search engines to access and retrieve these digital data. There is a trend to increase the volume of data a search engine can access and index. This has obvious advantages, but produces new challenges. One needs to increase retrieval specificity while maintaining an acceptable sensitivity.

The theory of signal detectability shows that even for very modest retrieval mechanisms, it is quite possible to move along the receiver-operating-characteristic curve to areas of very high specificity. The problem is the price one has to pay. Assuming a constant overall performance, increasing specificity will decrease sensitivity, thus causing the query to miss some of the truly relevant records.

A solution is to modify the search engine, such that it can attain higher specificity without sacrificing sensitivity. In other words, the overall performance of the search has to increase. Alternatively, one may estimate a relevance score for each record, and then sort the results descendingly.

In this paper we present a data retrieval system, AnonMed, which implements a sentence-level search engine and defines a relevance metric on the MEDLINE/PubMed data. We explain advantages and potential caveats, and demonstrate its capabilities through a few real examples.

Results

Almost all of the articles AnonMed (AM) retrieves overlap with that of PubMed (PM) (98.9%). However the first articles in AM are mainly the relevant ones while in PM relevant articles are mixed with false positives across the results. In example one, the true positive rate (TPR) of AM was 90%, but only 52% in PM. And in example two TPR was 83.3% vs 37.5%, respectively.

Conclusions

By using a sentence level matching, AnonMed is able to deliver higher specificity, thus eliminating false positive articles. Also, by introducing relevance metric, the most relevant articles are shown first, where the user focuses most. Furthermore, by composing the matching sentences and highlighting the keywords, AnonMed shrinks the displayed text, and hence the time the user spends for the 'scan & eliminate' process.

Background / Statement of problem

Nowadays a substantial portion of the large amounts of data produced in different organizations is recorded in digital form. This enables search engines to access and retrieve these digital data. There is a trend to increase the volume of data a search engine can access and index. This has obvious advantages, but produces new challenges. One needs to increase retrieval specificity while maintaining an acceptable sensitivity. Specificity is the percentage of irrelevant records that can be eliminated, while sensitivity is the percentage of relevant records that can be found and shown to the user. For example, table 1 shows a scenario where there are a total of 16 million records indexed, among which there are 500 records relevant to a user's query. The query eliminates 15,997,900 of the 15,999,500 irrelevant records, hence a specificity of 99.99%. However, due to the large volume of records indexed, the tiny 0.01% fraction (100 – 99.99) will make 1600 records. On the other hand, assuming a sensitivity of 99%, 495 of the 500 relevant articles are located and retrieved. Hence there will be a total of 1600+495=2095 records returned to the user. However, the majority (1600/2095 = 76%) are false positives (FP), records that the search engine considers as relevant and shows to the user but in truth they are irrelevant, and should have been eliminated. This may not be acceptable from the user's perspective. Therefore the 99.99% specificity is not sufficient in this case. One would like to operate with a higher specificity. This is a common situation, given the expansion rate of digital repositories, and the growing coverage of search engines. The need to operate with very high specificity is mainly caused by the big number of irrelevant records for any given user query (UQ).

The theory of signal detectability shows that even for very modest retrieval mechanisms, it is quite possible to move along the receiver-operating-characteristic (ROC) curve to areas of very high specificity [1, 2]. The problem is the price one has to pay. Assuming a constant overall performance (measured by odds ratio (OR), for instance), increasing specificity will decrease sensitivity, thus causing the query to miss some of the truly relevant records. Continuing with the example of table 1, increasing specificity to 99.9999% will decrease false hits to 16, a more acceptable count. However, the sensitivity decreases from 99% to 50%. This means 250 out of 500 truly relevant articles won't be retrieved for the user, as shown in table 2. A solution is to modify the search engine, such that it can attain higher specificity without sacrificing sensitivity. In other words, the overall performance of the search has to increase. Alternatively, one may estimate a relevance score for each record, and then sort the results descendingly. This way, majority of FP records will be pushed down the list of results.

PubMed is the online service from National Library of Medicine (NLM) that indexes over 16 million published biomedical articles. A major method in PubMed to retrieve relevant articles is using a controlled vocabulary called Medical Subject Headings (MeSH). MeSH is used for indexing articles, and provides a consistent way to retrieve information that may use different terminology for the same concepts [3]. By design MeSH is optimal for finding articles that are about a biomedical concept that has been included in the MeSH vocabulary. Also, one can use the system to find articles that are about 'concept one' and simultaneously are about 'concept two', a Boolean 'and' operation. However, we observe that when a user searches PubMed for two (or more) concepts, usually the user is implicitly expressing interest in articles that talk about some

potential “relationship” between the concepts, rather than retrieving articles that talk about each concept separately. MeSH is not designed for this scenario, and consequently may not attain high specificity. In reality, the system retrieves articles that are indexed under each of the concepts, and leaves the rest of the work to the user. Then the user has to go through a multistep ‘scan and eliminate’ process, where he reads the titles (or quickly scans the abstracts), and decides whether to eliminate the article or leave it for the next round of more in-depth screening.

For example, given a query like “antidepressant AND suicide”, any article that contains these two terms is retrieved by PubMed to the user. Included are articles that use the terms ‘antidepressant’ and ‘suicide’ in completely separate sections/paragraphs without any relation claimed between them. However, if the user’s question is the relation between those concepts (as often it is when the user uses the AND operator), these “positive” articles may not be relevant. The article-level co-occurrence assumption of PubMed makes the retrieval process fairly sensitive, but it would become nonspecific at the same time.

We believe co-occurrence of terms at the ‘sentence-level’ is a more specific retrieval method than the co-occurrence at ‘article-level’. If one could find articles where the terms co-occur in the same sentence (or adjacent sentences), it would be a better indication of potential relationship between the terms discussed in the article (compared to the case where the terms co-occur in the article each in separate and distant sentences).

In this paper we present a data retrieval system, AnonMed, which implements a sentence-level search engine and defines a relevance metric on the MEDLINE/PubMed data. We explain advantages and potential caveats, and demonstrate its capabilities through a few real examples.

Implementation

To implement the sentence-level search engine, through a lease contract with National Library of Medicine, we obtained MEDLINE data in extensible markup language (XML) format. We designed and implemented algorithms to extract title, abstract, and citation information from each XML article record, then scanned through the abstract text to detect and separate sentences. To detect a sentence we used ‘.’, ‘?’, and ‘!’ as delimiters. We then joined back consecutive sentences where the period was sandwiched by single capital letters, some specific words such as ‘etc.’ and ‘et. al’, or by digits such as in ‘0.05’.

We designed a database with 2 tables, to load the sentences. Table 3 shows the fields and their definitions. Table 1 of the database contains the sentences, the bulk of data, where an index is created for them. Field PMID is a unique integer number assigned by NLM to each article. Here we used PMID to link table 1 to table 2. Field SNTNCID is equal to 1 for article title, and then 2 and bigger for abstract sentences. Table 2 contains the citation information for each NLM article. There is a many-to-one relationship between table 1 and table 2. Table 1 is used to match user query (UQ) to indexed articles, while Table 2 is used to retrieve citation information for a given PMID.

We designed and implemented an application to receive UQ, prepare the query in SQL language, interrogate the database, format the database results in HTML language, and post it back to the user's browser. The UQ can simply be composed of one or a few words, separated by space. By default the system ANDs the words. Also, Boolean operators OR and NOT are supported. One can use asterisk * for truncation, parentheses () for grouping, and quotes "" for exact phrase matching. These are in accordance with PubMed query language.

The system writes all the sentences matching the query in an HTML report, where the matched keywords are highlighted. The publication information for the article where the sentence was found is then added, as well as a hyperlink such that the user can easily navigate to the respective PubMed (PM) article, for potential drill down and more in-depth evaluation, Figure 1.

We used freely available open source software to build the search engine. We used Perl to pre-process data, and to write the query application [4], MySQL to implement the database [5], and Apache to serve user's HTTP requests [6]. We installed our server with Fedora operating system [7], hence the so-called LAMP architecture. We used XHTML to produce the user interface and the reports [8].

Relevance metric

Given an article record, with title (one sentence), 9 abstract sentences (on average), and MeSH terms (concatenated together and treated as one sentence), one can assign importance weights to each of the three sentence types (title, abstract, MeSH). Then one can combine the types to define several levels of 'relevance'. Thus one tries to measure how closely an article answers user's query. Then one can sort the returned results by the relevance metric. This pushes the most relevant articles to the top of the results list. Therefore the user would see the most relevant results first. This is a useful and time-saving feature.

Table 4 defines eight relevance levels, hence a discrete metric. Assuming user's query is 'word1 word2', in relevance level one, both the words should appear in title, and both words should appear in at least one sentence in abstract, and both words should appear in the MeSH terms, a stringent set of criteria. This we believe would indicate, in majority of times, that the matched article would be of high relevance to the user's query, hence the first relevance level. The next levels are defined similarly, only the combinations of the types of sentences are different. Level 8 is different from the rest, as we first concatenate together all the sentences of an article, including title, all abstract sentences, and all the MeSH words. This makes one big 'sentence' from the whole article, where user's query is matched against. For example, word1 can be in title, while word2 in MeSH words or in any of the abstract sentences. This level adds to the sensitivity of the search engine, thus reducing probability of missing any relevant article. However level 8 has a low specificity, which is the reason we assigned the lowest relevance level to it.

Evaluation method

We designed a 'randomized and blinded' study to evaluate the AnonMed search engine, and compare it to PubMed. The intention was to decrease evaluation bias as much as possible. We tested two hypotheses: H1. The overall sensitivity and specificity of

AnonMed (AM) is similar to PubMed (PM). In other words, given a user query, the collection of articles returned by AM is almost the same as PM. H2. The most relevant articles are listed at the start of the AM results, while in PM they are mixed with false positive articles along the whole list. We started with a user query (UQ). We chose a pre-defined article count n , like 10. We queried AM with UQ, and saved PMIDs of first n articles within each relevance level, hence a total of $8n$ PMIDs. Likewise we queried PM with UQ, and saved the first $8n$ PMIDs. Then we wrote a program to which we fed the two lists of $8n$ PMIDs. The program made a unique list of PMIDs, and then put them in random order, using a random number generator. Then the program queried the database for each PMID in the randomized list, and wrote an HTML report where the article contents (title, abstract, and MeSH) are ordered according to the randomization list. Keywords were highlighted in the HTML report, to facilitate evaluation process. Then two biomedical experts (MSS and JS) inspected the articles independently, and assigned true positive (TP) or FP labels to each, thus defining the ‘gold standard’. To resolve potential discordance between the two raters, a discussion was made on each of the discordant articles to reach a consensus. Then the program transferred the TP and FP assignments back to the query results of each of the PM and AM, thus ‘breaking the blind’. Finally we estimated the true positive rate ($TPR = 1 - FPR$) for each of the relevance levels of AM, and consecutive bins of size n in PM.

To analyze the FPR data, and to attach statistical significance, we used local regression implemented in package ‘locfit’ of R statistical language [9, 10]. Also, to measure inter-rater agreement, we used Cohen's kappa (it measures the agreement between the evaluations of two raters when both are rating the same object. A value of 1 indicates perfect agreement. A value of 0 indicates that agreement is no better than chance.)

Results

Example 1: Role of ‘infection’ in ‘sudden infant death syndrome’ (SIDS)

SIDS is death of an infant less than one year old that cannot be explained after thorough medical investigation [11]. Despite years of research, no definitive cause has been found, but there are many potential factors proposed by investigators, such as position of baby during sleep, use of pacifier, parents’ smoking, infection, change in temperature, etc. In this example the user wants to retrieve existing literature on SIDS that might point to infection as a potential cause of death in SIDS (or explains absence of such causal relation).

We used the query

(sids or “sudden infant death”) and (infection or infections or infectious or “communicable diseases”)

on both PM and AM. PM returned 697 articles (as of 20Feb2006), while AM returned 666, where 659 of them were listed in the PM results (98.9%).

Table 5 shows count of articles in each AM relevance level. We used a cutoff of $n = 10$ to compose the randomization list. For levels where the total returned articles were smaller than 10, we used all available. This made a list of 69 PMIDs. Then we added the first 69

articles from PM, thus making a list of 138 PMIDs. The TPRs were estimated by the method explained in the Evaluation section. The inter-rater agreement was 94% (7 discordant articles among the 115 unique PMIDs), with Kappa of 0.873 (p-value < 0.001).

Figure 2 shows the TPR (the red dots) in the 8 groups of PMIDs per search engine. We fitted smoother curve (solid blue lines) to the observed binary data (TP versus FP), to facilitate visualizing the trend. Also, we estimated 95% global confidence bands (the dashed black curves), for inference. There is a decreasing TPR trend in AnonMed. However, the trend in PubMed is not monotone. The average TPR in the first 69 articles of PubMed was 52%, while the estimated average TPR for the first 69 articles of AnonMed was 90%.

Table 6 shows example of an FP article. All instances of the keywords in the article are highlighted and shown. Both infection and SIDS are mentioned in two separate sentences of abstract. Plus both of them are in the MeSH terms. However, no relation between the two is declared.

Example 2: finding ‘questionnaires’ for measuring ‘health literacy’

Health literacy (HL) is the degree to which individuals have the capacity to obtain, process, and understand basic health information and services needed to make appropriate health decisions [12]. In this example the user has a research project where he wants to measure HL. He is interested in finding papers that give clues about existing questionnaires/instruments for HL.

We used the query

"health literacy" AND (instrument OR question* OR measur* OR scal* OR assessment* OR index* OR test*)*

and PM returned 150 articles, while AM returned 130 where 125 of them were shared with PM (96.1% overlap). There were 4 articles in level 2, 53 in level 6, and 73 in level 8. The inter-rater agreement was 83.7% (7 discordant articles out of 43), with a statistically significant (p-value < 0.001) Kappa of 0.68. The TPR in PM for the first 24 articles was 37.5%, while TPR for the first 24 articles of AM was estimated at 83.3%.

In Figure 3 the decreasing trend of TPR in AM is apparent, while that of the PM is almost horizontal. Note the TPR starts from a much higher point in AM compared to PM. Also, the maximum TPR attained in PM is much lower.

Discussion

AnonMed

It seems the bigger portion of queries sent to MEDLINE/PubMed are multi-word queries, where two or more concepts are included in the query (as opposed to single-concept queries). For queries with two (or more) concepts, usually the user is looking for articles that are about 1.concept one, 2.concept two, and 3.some relationship between the concepts. Currently PubMed retrieval mechanisms satisfy objectives 1 and 2 better than 3. Drawing on linguistics, the chance of the article claiming some relation between the two concepts

is higher when they co-occur within a sentence than an article (or abstract). This is the basis for creating AnonMed search engine.

There is a limitation on the amount of text a user is willing or able to scan. By using a sentence level matching, AnonMed (AM) is able to deliver higher specificity, thus reducing false positive (FP) articles. Also, by introducing relevance metric, the most useful articles are shown first, where the user focuses most. Furthermore, by composing the matching sentences and highlighting the keywords, AM shrinks the text and the time the user spends for the ‘scan & eliminate’ process. The two examples explained in the paper demonstrated that the maximum TPR is attained at the start of results in AM, but not in PM.

Assuming AM returns almost the same collection of articles as PM, a question is the location of the false positive articles in the AM results. We believe relevance levels 7 and 8 would contain majority of FPs. Level 6 may also contains FPs, as the presence of relation detected by sentence-level co-occurrence is not supported by MeSH that would have shown the article is about the matched words.

The collection of articles retrieved by AM is not exactly the same as PM. We see a couple of reasons: 1.query mapping has not been implemented in AM at the time of this evaluation. This means even though we used the same UQ in AM and PM, PM usually maps the UQ to something that can be different. This is why we used a query like ‘sids or “sudden infant death”’, to mimic the query mapping of PM. 2.For this study, we used the 2006 base distribution of MEDLINE. This means PM contains about 4 months of new articles (approximately 200 thousand articles) that were not uploaded to AM yet.

One can classify users of PubMed into two sub-populations: 1.users with expert knowledge of the PM query language, 2.users who know little about retrieval features and query language of PM. We believe majority of PM users are of the second group. The example presented in Table 1 is mostly fitting the expert user, as it has an OR of one million. An OR of one million means, given one million pairs of articles where one is relevant and one is irrelevant, the query can classify all of the one million pairs correctly except one. Usually a non-expert user does not formulate a query with such high performance. In other words, for the second bigger population of non-expert users there is a more serious need for improvement of the search engine. Furthermore, we observe that multiword queries to PubMed may be more common than single words. This emphasizes the need to find articles that are not simply about each word, but articles that in addition declare relation between the words.

Table 2, from the user’s perspective, seems desirable: 1.the count of articles returned by the search engine is more manageable, 266 versus the previous count of 2095 in Table1; 2.the percentage of TP records is much better/higher, 94% versus the 24% of Table 1. However, the user can not realize that a large number of relevant records are not retrieved by the search engines, 250 out of 500, hence lost.

In defining the relevance levels, we gave lower weight to MeSH compared to title or abstract. If query words co-occur in MeSH, it is no indication of the article claiming any sort of relation between them. However we observe that if a term appears in MeSH, then the article is about the term and its concept. Hence we think combining title or abstract sentences with MeSH gives a good relevance, as in the earlier relevance levels. We note

that relevance level 8 is similar to what PubMed does, the article-level co-occurrence method.

Given a fixed specificity, the FPR estimate of 76% in Table 1 would change based on the arbitrary assumptions of 1.number of relevant articles for a given query and 2.query sensitivity. However, one would see that there is a practical ceiling for the number of relevant articles a user is willing to read. If a query returns too many relevant articles, usually the user will narrow down to reach a number that is plausible to study further. We believe a ceiling of 1000 is quite an over-estimation. Given that, the FPR would be $1600/(1600+990) = 61.8\%$, and still has much room to improve. Also, we believe sensitivity of 99% is already an over-estimation. These two show that FPR of 62% can only be a conservative under-estimation.

We note since AM matches UQ against each single sentence, having too many words in UQ might return no articles in the first few relevance levels.

PubMed

One may try to use some of the PubMed features to better answer the multi-word queries. Three methods come to mind: 1.Searching free text (includes title and abstract) to find articles where the two concepts ‘co-occur’. However, there may be many articles where the concepts are mentioned in completely separate sentences or sections with no indication of any sort of relation between them, hence false positive results. 2.One can limit the search to the titles only. Then if the (two) concepts appear in the title, it has a high probability that some sort of relation is declared between them in the article. Although this method could attain fairly high specificity, it may miss relevant articles as it does not utilize any of the sentences of the abstract, hence a potentially low sensitivity. 3.If the two concepts can be expressed in consecutive words, one may be able to use quoting (the operator “”), so that to instruct PubMed to retrieve articles where the words appear exactly (in the same proximity and order) as they are in the quoted phrase. However, this does not seem to be a common case though.

If the two or more concepts the user is asking, have hierarchical relation in the MeSH, then MeSH can show high specificity. For example, when the user is interested in adverse effects of antidepressant therapy, the MeSH subheading ‘adverse effects’ to the MeSH heading ‘antidepressive agents’ is a good query.

A weakness of using MeSH vocabulary is ‘time to indexing’, for new concepts. For example many concepts in bioinformatics are still absent in MeSH. Note the two concepts of example 1 and one of the two concepts in example 2 are the terms that are indexed by MeSH. Retrieval performance of PM for terms not indexed may be even lower.

For future

There are topics one would like to cover in a search engine like AM. Some are: 1.To implement a sentence-level search engine, one needs to detect sentences in a text. Assuming one would be using computers to detect and separate sentences in the abstract texts, ‘period’ is the most common sentence delimiter. However, it is used for other purposes within a sentence. Examples are decimal points of numbers, and periods in acronyms and abbreviated words. One needs to improve on the algorithms detecting

sentences. This is a natural language processing problem. 2. One needs to add algorithms to AM such that it can map UQ to MeSH terms automatically. Features like ‘stemming’, syntax checking and suggesting, would be useful too. 3. To build relevance metric with finer granularity, one may define an adjacency metric, where distance of keywords in the sentence or adjacent sentences are measured and taken into account. 4. To increase specificity, one can implement NLP algorithms to detect and verify different types of relationships declared in a sentence between the words, by detecting and parsing verbs and other syntactical parts. 5. There is a movement toward open access full text papers. Indexing full text papers would be a significant improvement. 6. It is useful to have field-specific search capability, as implemented in PM via operator []. In designing these and other new features, one would prefer to incorporate only features that can be implemented in fast algorithms such that the total response time is short. Therefore the interactive nature of the search engine is preserved.

It is important to improve and add more features to the search engine. However, we believe it is equally important to implement algorithms such that the search engine utilizes the optimal features automatically. In other words, the default setting of the search engine should be such that for the majority of users, who are non-expert users, the system is already utilizing the advanced features. This is in contrast with the approach that makes the features more user-friendly, and then hopes that the non-expert user would utilize the advanced search capabilities (over the simple less efficient ‘default’ setting), Figure 4.

Tables

Table 1. Query with specificity of 99.99% is insufficient for a database of 16 million records.

		The truth		
		relevant records	irrelevant records	
User's query	records returned to user	495	1,600	2,095
	records eliminated	5	15,997,900	
		500	15,999,500	16,000,000

given OR 1,000,000.00
 given specificity 99.99%
 sensitivity 99.01%
 FPR 76.37%

Table 2. The price for a very high specificity: Missing a large number of relevant records.

		The truth		
		relevant records	irrelevant records	
User's query	records returned to user	250	16	266
	records eliminated	250	15,999,484	
		500	15,999,500	16,000,000

given OR 1,000,000.00
 given specificity 100.00%
 sensitivity 50.00%
 FPR 6.01%

Table 3. Database tables, and their fields

Database table1		
Field	Description	Indexed
PMID	PubMed ID number	no
SNTNCID	sentnece ID number	no
Sentence	text of the sentence	yes

Database table2		
Field	Description	indexed
PMID	PubMed ID number	yes
Citation	Citation information for the article	no

Table 4. The eight relevance levels defined by AnonMed.

Relevance level	Query must match
1	T and A and M
2	T and A
3	T and M
4	A and M
5	T
6	A
7	M
8	TAM

T = title

A = at least one abstract sentence

M = concatenated MeSH terms

TAM = title, abstract, and MeSH
concatenated into one sentence

Table 5. Count of articles in each AM relevance level for query of Example 1.

Relevance	Count of articles
L1 T&A&M	31
L2 T&A	3
L3 T&M	24
L4 A&M	68
L5 T	6
L6 A	150
L7 M	205
L8 TAM	179
Total	666

Table 6. A false positive article for query of Example 1, where query words do co-occur, both in text and in MeSH.

DiFranza JR, Aligne CA, Weitzman M. **Prenatal and postnatal environmental tobacco smoke exposure and children's health.** *Pediatrics*. 2004 Apr;113(4 Suppl):1007-15.

... A large literature links both prenatal maternal smoking and children's ETS exposure to decreased lung growth and increased rates of respiratory tract **infections**, otitis media, and childhood asthma, with the severity of these problems increasing with increased exposure. **Sudden infant death** syndrome, behavioral problems, neurocognitive decrements, and increased rates of adolescent smoking also are associated with such exposures. ...

[MeSH] drug effects. etiology. adverse effects. Animals. Asthma. etiology. Child. Child Behavior. drug effects. Embryonic and Fetal Development. Female. Humans. **Infant**. Intelligence. drug effects. Otitis Media. etiology. Pregnancy. Respiratory Tract **Infections**. Smoking. adverse effects. **Sudden Infant Death**. etiology. Tobacco Smoke Pollution. analysis

Availability and requirements

Project name: AnonMed

Project home page: <http://www.anonmed.com>

Operating systems: Platform independent

Programming language: Perl

Other requirements: None

License: Free, anyone may use the service

Any restrictions to use by non-academics: None

List of abbreviations

AM: AnonMed

FP: False Positive

FPR: False Positive Rate

HTML: HyperText Markup Language

HTTP: HyperText Transfer Protocol

LAMP: Linux Apache MySQL Perl

MeSH: Medical Subject Headings

NLP: Natural Language Processing

OR: Odds Ratio

PM: PubMed

PMID: PubMed ID

ROC: Receiver Operating Characteristic

SIDS: Sudden Infant Death Syndrome

TP: True Positive

TPR: True Positive Rate

UQ: User Query

XHTML: eXtensible HyperText Markup Language

XML: eXtensible Markup Language

Competing interests

A patent application has been filed, by authors of this paper.

Authors' contributions

MSS conceived of the method, carried out its implementation, participated in its evaluation, and drafted the manuscript. JS participated in the evaluation and drafting the manuscript, and gave feedback to improve the search engine. WAK participated in drafting the manuscript, and gave feedback to improve the search engine. All authors read and approved the final manuscript.

Acknowledgements

None.

References

1. Peterson WW, Birdsall TG, Fox WC: **The theory of signal detectability.** *Transactions of the IRE professional group on information theory* 1954, **4**:171-212.
2. Tanner WP, Swets JA: **A decision-making theory of visual detection.** *Psychol Rev* 1954, **61**(6):401-409.
3. **Medical Subject Headings** [<http://www.nlm.nih.gov/mesh/meshhome.html>]
4. **Comprehensive Perl Archive Network** [<http://www.cpan.org>]
5. **MySQL AB** [<http://www.mysql.com/>]
6. **The Apache HTTP Server Project** [<http://httpd.apache.org/>]
7. **Fedora Project, sponsored by Red Hat** [<http://fedora.redhat.com/>]
8. **The Extensible HyperText Markup Language** [<http://www.w3.org/TR/xhtml1/>]
9. R Development Core Team: *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing; 2004.
10. Loader C: *Local Regression and Likelihood.* New York: Springer; 1999.
11. Willinger, M., James, L.S., and Catz, C: **Defining the Sudden Infant Death Syndrome (SIDS): Deliberations of an Expert Panel Convened by the National Institute of Child Health and Human Development.** *Pediatric Pathology* 1991, **11**:677-684.
12. U.S. Department of Health and Human Services: **Healthy People 2010: Understanding and Improving Health. 2nd ed.** Washington, DC: U.S. Government Printing Office; November 2000.

Figure legends

Figure 1. AnonMed search engine returns results where almost all of the top articles are true positive and relevant.

Figure 2. The true positive rate (TPR) in AnonMed is high and is maximum at the start of the results; but in PubMed TPR may reach its maximum anywhere.

Figure 3. AnonMed shows consistent maximum TPR at the start of the results; but PubMed has almost a constant and low TPR.

Figure 4. AnonMed accepts queries in PubMed language, without need for modification.

Matching query 'sudden infant death' (infections or infectious or infectious or "communicable diseases"), simultaneously against 1 title, and 2 each contents in abstract, and 1 MeSH (Relevance Level 1).

Running time: 1 to 31 (in 0.050278 seconds).

1. An EP, Gould S, Koring JW, Fleming EA. Role of respiratory viral infection in SIDS: detection of viral nucleic acid by *in situ* hybridization. *J Pediatr*. 1993 Dec;171(4):271-8. [[View PubMed record](#)]

Matches:

- [TITLE] Role of respiratory viral **infection** in **SIDS**: detection of viral nucleic acid by *in situ* hybridization.
- There is considerable evidence suggesting that respiratory viral **infection** is involved in the genesis of the **sudden infant death** syndrome (**SIDS**), with rates of about 28 per cent of **SIDS** victims compared to about 13 per cent of controls.
- [MeSH] microbiology complications: etiology complication isolation & purification. Adenovirus, Human isolation & purification. Age Factors. DNA, Viral analysis Female Humans. *In Situ* Hybridization. **Infant**, **Infant**, Newborn, Lung, Male. Parainfluenza Virus 2, Human. isolation & purification. RNA, Viral analysis Research Support, Non-U.S. Gov't. Respiratory Syncytial Virus: isolation & purification. Respiratory Tract **Infections**: Season. **Sudden Infant Death**. **Virus Diseases**. **Virus**

Content:

- [Title] Role of respiratory viral **infection** in **SIDS**: detection of viral nucleic acid by *in situ* hybridization.
- [Abstract] There is considerable evidence suggesting that respiratory viral **infection** is involved in the genesis of the **sudden infant death** syndrome (**SIDS**), with rates of about 28 per cent of **SIDS** victims compared to about 13 per cent of controls. Since the techniques used previously are prone to under-reporting from autopsy material, non-otopic *in situ* hybridization (NISH) has been used to detect viral nucleic acid in lung in **SIDS**. Forty-five **SIDS** cases (30 males) were examined (age range 3 weeks-14 months, mean age 3.9 months). Thirty non-**SIDS** cases (15 males) were also examined (age range 3 weeks-24 months, mean age 9.0 months). Eleven of 40 (24.4 per cent) **SIDS** cases were positive by NISH, compared to 3 of 30 (3.3 per cent) non-**SIDS** cases ($P = 0.012$). There were eight cases of adenovirus type 3, two cases of respiratory syncytial virus (RSV), and one case of parainfluenza virus type 2. The one positive control case was adenovirus type 3. Only lung parenchyma was examined here. A detailed examination of the upper respiratory tract may increase the number of positive cases.
- [MeSH] microbiology complications: etiology complication isolation & purification. Adenovirus, Human isolation & purification. Age Factors. DNA, Viral analysis Female Humans. *In Situ* Hybridization. **Infant**, **Infant**, Newborn, Lung, Male. Parainfluenza Virus 2, Human. isolation & purification. RNA, Viral analysis Research Support, Non-U.S. Gov't. Respiratory Syncytial Virus: isolation & purification. Respiratory Tract **Infections**: Season. **Sudden Infant Death**. **Virus Diseases**. **Virus**

Figure 1

2. Vega A, Chen J, Spector SA, Spector CD, Higgins TC. Viral and human hypermethylation levels in SIDS and infectious death. *Acta Paediatr*. 1994 Jun;83(6):624-9. [[View PubMed record](#)]

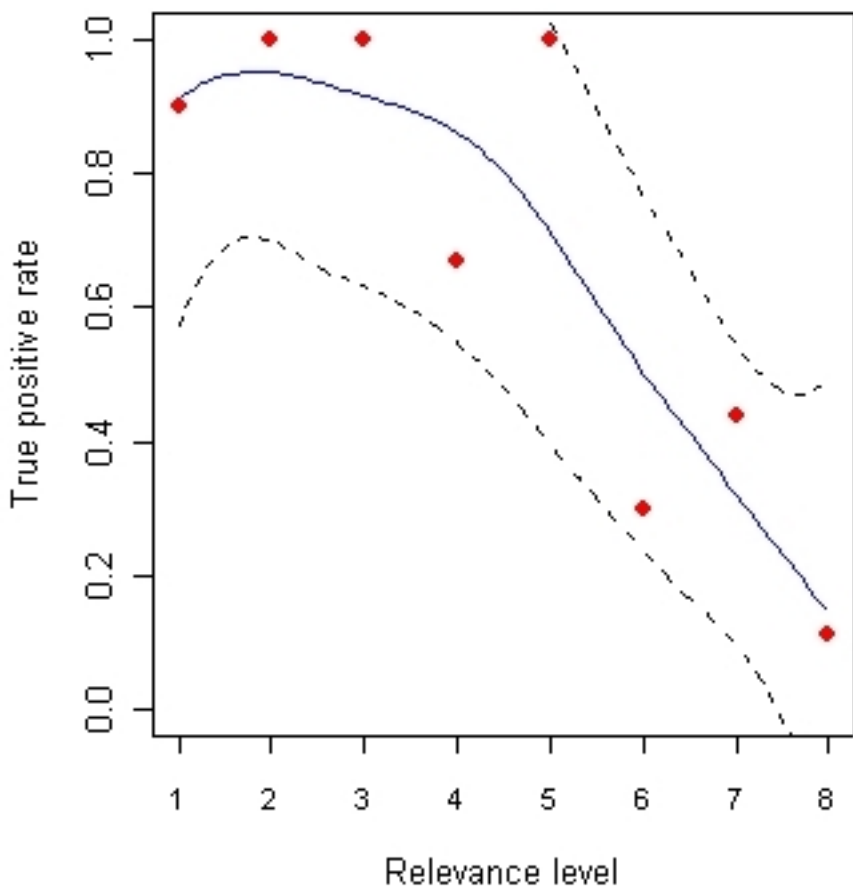
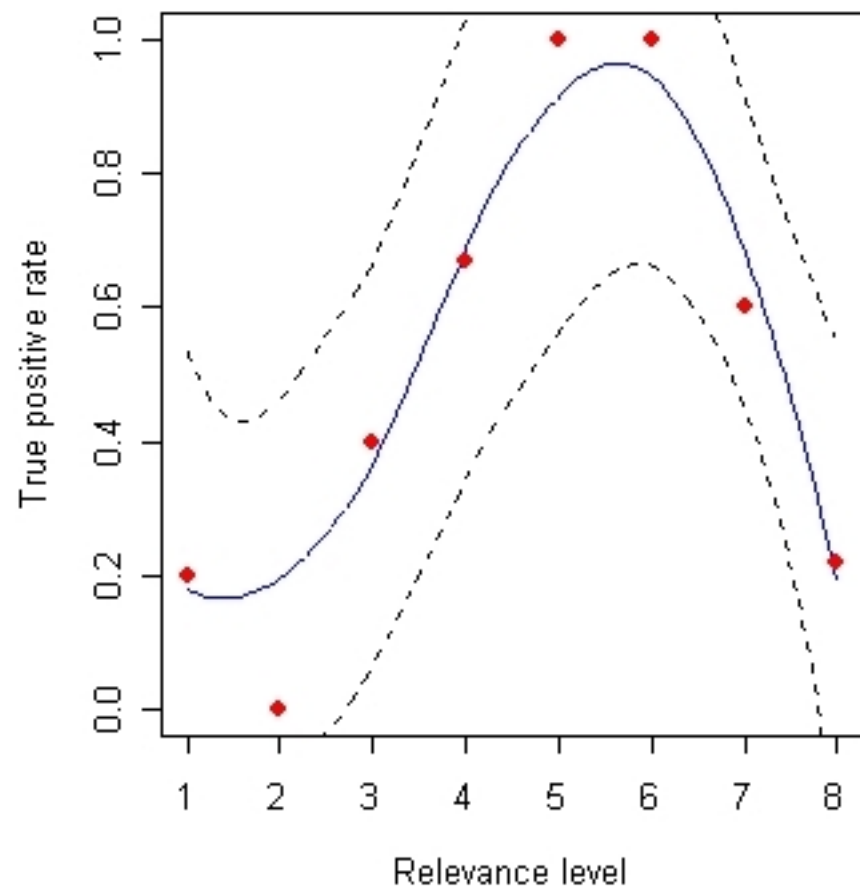
AnonMed**PubMed**

Figure 2

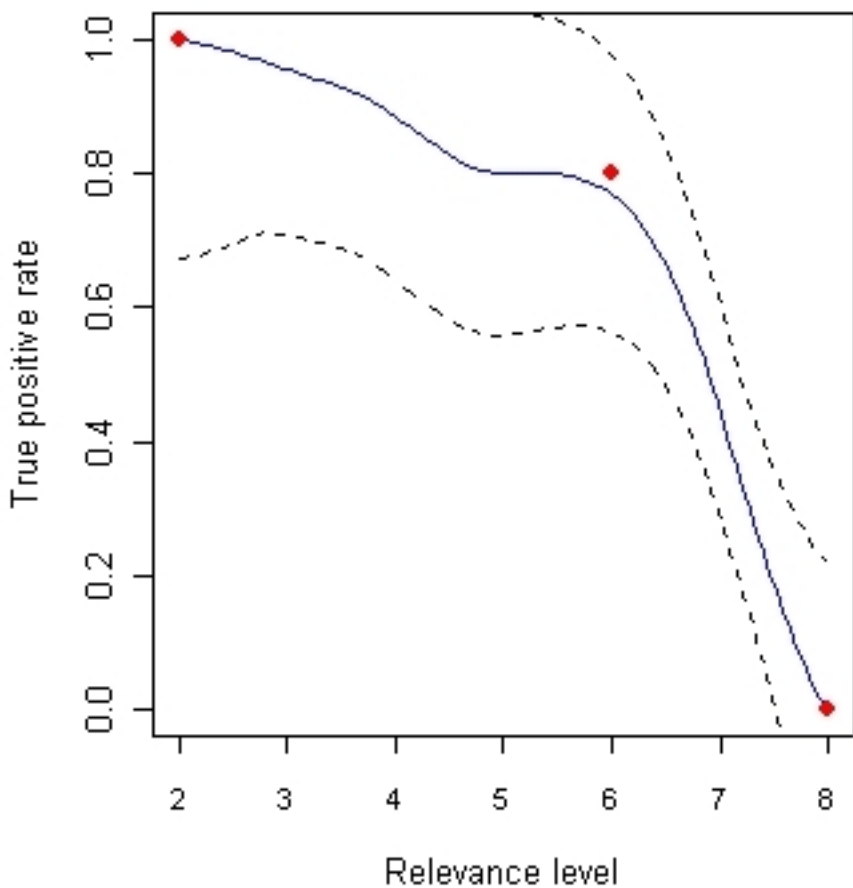
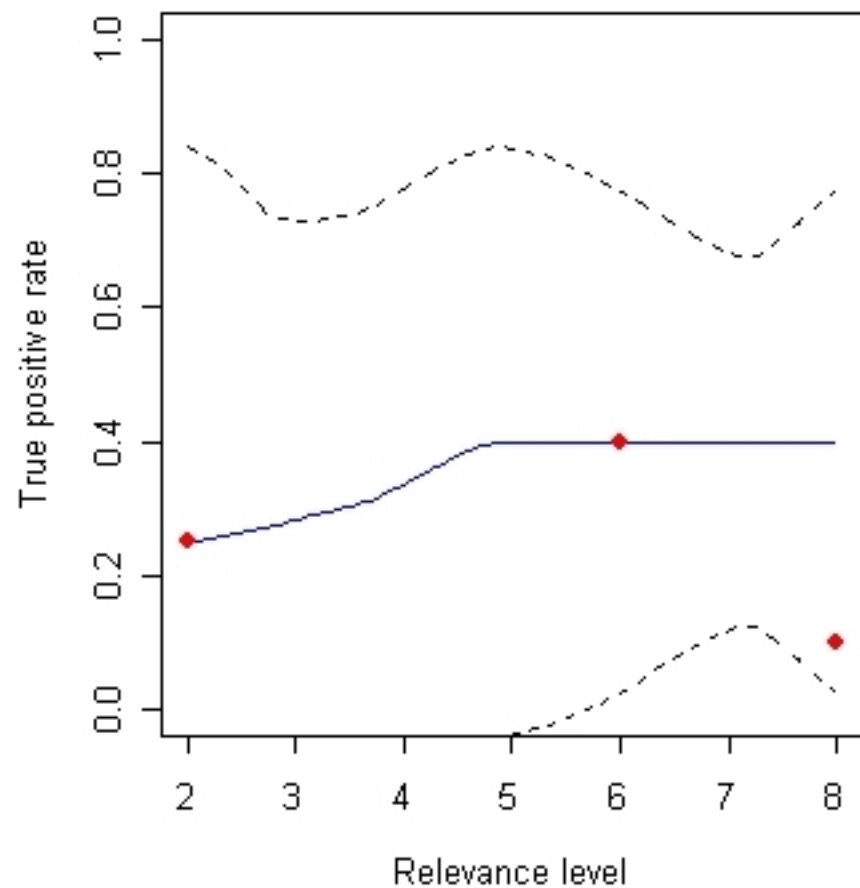
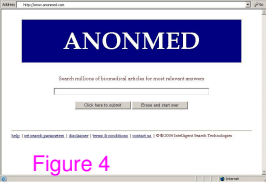
AnonMed**PubMed**

Figure 3



ANONMED

Search millions of biomedical articles for most relevant answers

Click here to submit

Erase and start over

[Help](#) | [refman¶meters](#) | [disclaimer](#) | [terms&conditions](#) | [contactUs](#) | © 2008 Intelligent Search Technologies

Figure 4