



Text categorization models for identifying unproven cancer treatments on the web

Yin Aphinyanaphongs PhD
Constantin Aliferis MD/ PhD
Department of Biomedical Informatics
Vanderbilt University
Medinfo 2007

Motivation

- “The killing of all parasites and their larval stages together with removal of isopropyl alcohol and carcinogens from the patients’ lifestyle results in remarkable recovery (from cancer) generally noticeable in less than a week”
- Internet allows quacks to advocate unproven, and inaccurate treatments.

Why now?

- 65% of cancer patients searched unproven treatments.
- 12% pursued unconventional therapies online.
- 83% of cancer patients use at least one unproven treatment.
- Patients are ill-equipped to evaluate treatments.

Metz JM, et al. 2003.
Richardson MA, et al. 2000.
Sagaram S, et al. 2002.
Ernst E, et al. 2002.

Costs of Quack Treatments

- Financial costs.
- Psychological costs.
- Delayed application of proven modalities.
- Documented fatal, adverse outcomes.

Hainer MI, et al. 2000.
See KA, et al. 2006.
Bromley J, et al. 2005.
Mularski RA, et al. 2006.

Previous Approach - Self Regulation

- Self regulation.
 - Health on Net Foundation.
 - Limitations
 - Requires knowledge of certification.
 - Vigilant public to report violations.

Health on the Net; <http://www.hon.ch/>
Eysenbach G, et al. 2002.

Previous Approach - Expert Rating Tools

- Expert rating tools.
 - Limitations
 - Require knowledgeable public to apply.
 - Difficult to validate & most instruments not validated.
 - Time consuming to produce.
 - Do not produce consistent ratings.
 - Ratings tools are not appropriate for use on complementary/ alternative medicine sites.
 - Manual review requires reviewer time and limits in web sites selected to review.

Bernstam EV, Shelton DM, et al. 2005.

Kim P, et al. 1999.

Bernstam EV, Sagaram S, 2005.

Ademiluyi G, et al. 2003.

Walji M, et al. 2004.

Ideal Solution

- Validated.
- Easy to use by health care consumers.
- Works on any webpage.

Hypothesis

- Automated classification approaches effectively identify web pages that make unproven claims.

Quackwatch.org

- 36 year old nonprofit organization.
- “combat health related frauds, myths, fads, fallacies, and misconduct”
- 152 person scientific and advisory board.

Gold standard

- ▶ Design: positives drawn from quackwatch.org and negative controls from the Web.
- ▶ Considered web pages about 8 quack treatments randomly selected from quackwatch.org.
 - ▶ “Cure for all Cancers”
 - ▶ “Mistletoe”
 - ▶ “Krebiozen”
 - ▶ “Metabolic Therapy”
 - ▶ “Cellular Health”
 - ▶ “ICTH”
 - ▶ “Macrobiotic Diet”
 - ▶ “Insulin Potentiation Therapy”
- ▶ Applied to Google appending “cancer” and “treatment.”
- ▶ Top 30 results for each treatment labeled by the authors.
- ▶ Resulted in 240 web pages.

Corpus Labels

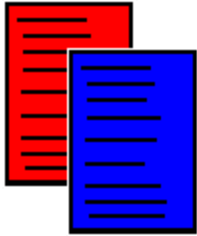
- Excluded
 - not found (404 response code) error pages
 - no content pages
 - non-English pages
 - password-protected pages
 - pdf pages
 - redirect pages
 - pages where the actual treatment text does not appear in the document
- The authors labeled 191 out of 240 web pages as making unproven claims or not (Inter-rater Reliability - Kappa 0.76)
- 93 positive web pages/ 98 negative web pages

Content Representation

- Bag of Words.
 - Removed all content between style and script tags.
 - All tags (including style and script tags removed).
 - Replaced all punctuation with spaces.
 - Split on spaces to obtain words.
 - Stemmed the words.
 - Applied stop word list.
 - Removed any words appearing less than 3 pages.
 - Encoded words using log frequency with redundancy scheme.

Basic Framework

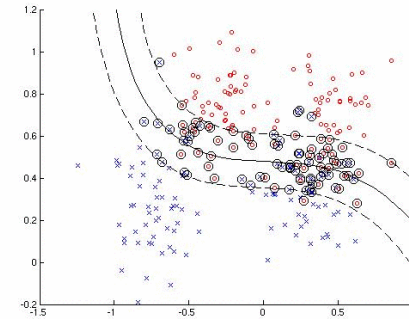
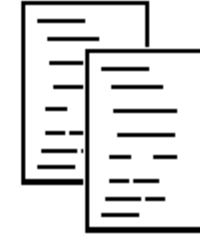
Labeled
Web Pages



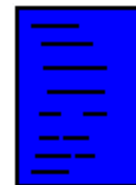
```
switch (iFilterType)
{
case CM_FILTERHIGHPASS:
case CM_FILTERBANDPASS:
    hrng[2] = CreateEllipticRgn(x11, y11, x12, y12);
    break;
case CM_FILTERLOWPASS:
case CM_FILTERBANDPASS:
    hrng[0] = CreateEllipticRgn(x11, y11, x12, y12);
    hrng[1] = CreateRectRgn(0, 0, iW, iH);
    hrng[2] = CreateRectRgn(0, 0, iW, iH);
    CombineRgn(hrng[2], hrng[0], hrng[1], RGN_XOR);
    DeleteObject(hrng[0]);
    DeleteObject(hrng[1]);
case CM_FILTERBANDPASS:
    hrng[0] = CreateEllipticRgn(x21, y21, x22, y22);
    hrng[1] = CreateEllipticRgn(x11, y11, x12, y12);
    hrng[2] = CreateRectRgn(0, 0, iW, iH);
    CombineRgn(hrng[2], hrng[0], hrng[1], RGN_XOR);
    DeleteObject(hrng[0]);
    DeleteObject(hrng[1]);
case CM_FILTERBANDPASS:
    hrng[0] = CreateRectRgn(0, 0, iW, iH);
    hrng[1] = CreateRectRgn(0, 0, iW, iH);
    hrng[2] = CreateRectRgn(0, 0, iW, iH);
    CombineRgn(hrng[2], hrng[0], hrng[1], RGN_XOR);
    DeleteObject(hrng[0]);
    DeleteObject(hrng[1]);
    hrng[0]=CreateRectRgn(0, 0, iW, iH);
    hrng[1]=CreateRectRgn(0, 0, iW, iH);
    CombineRgn(hrng[2], hrng[0], hrng[1], RGN_AND);
    // Tegnet det endelige carødet i rødt
    FillRgn(PaintInfo.hdc, hrng[3], hbrRed);
    // Fjern de allekerte regionene, de er bare midlertidige
    for(i=0; i<4; i++)
        if (hrng[i]!=NULL)
            DeleteObject(hrng[i]);
}
```

Ranking by:
Machine Learning
Quackometer
Google PageRank

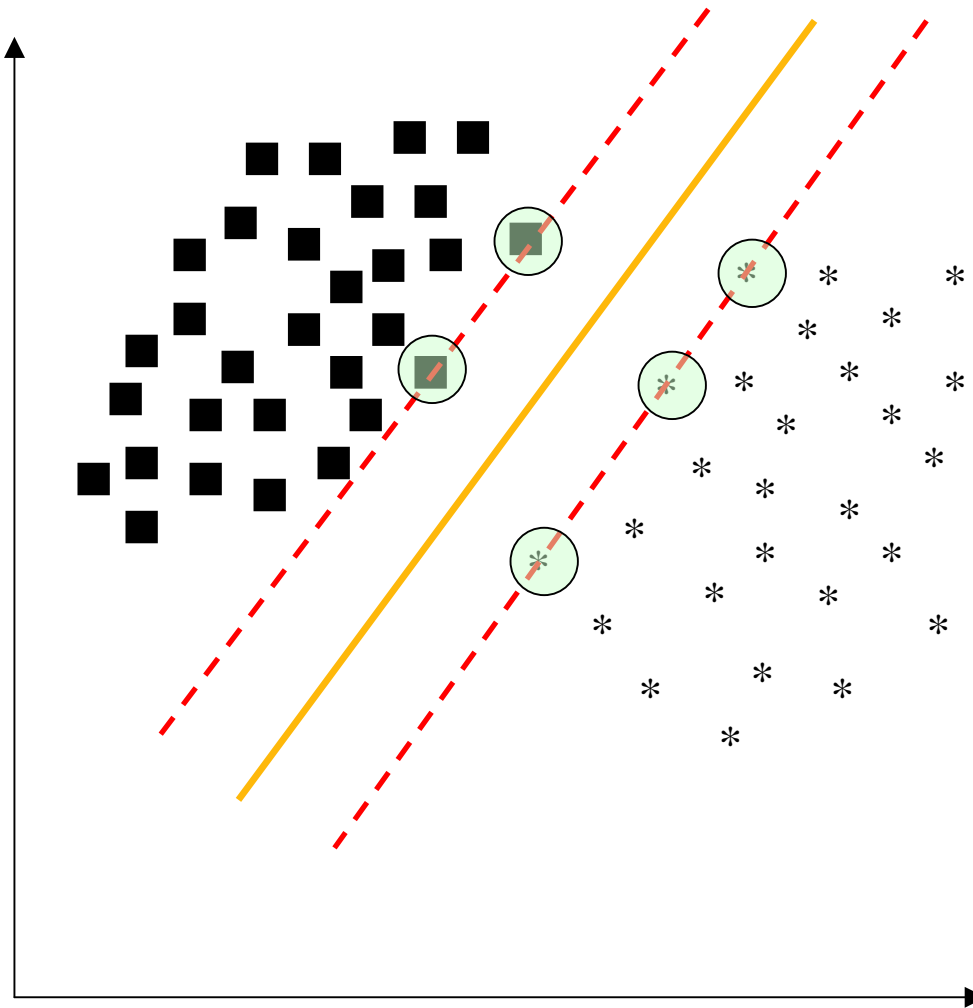
Unseen Web
Pages



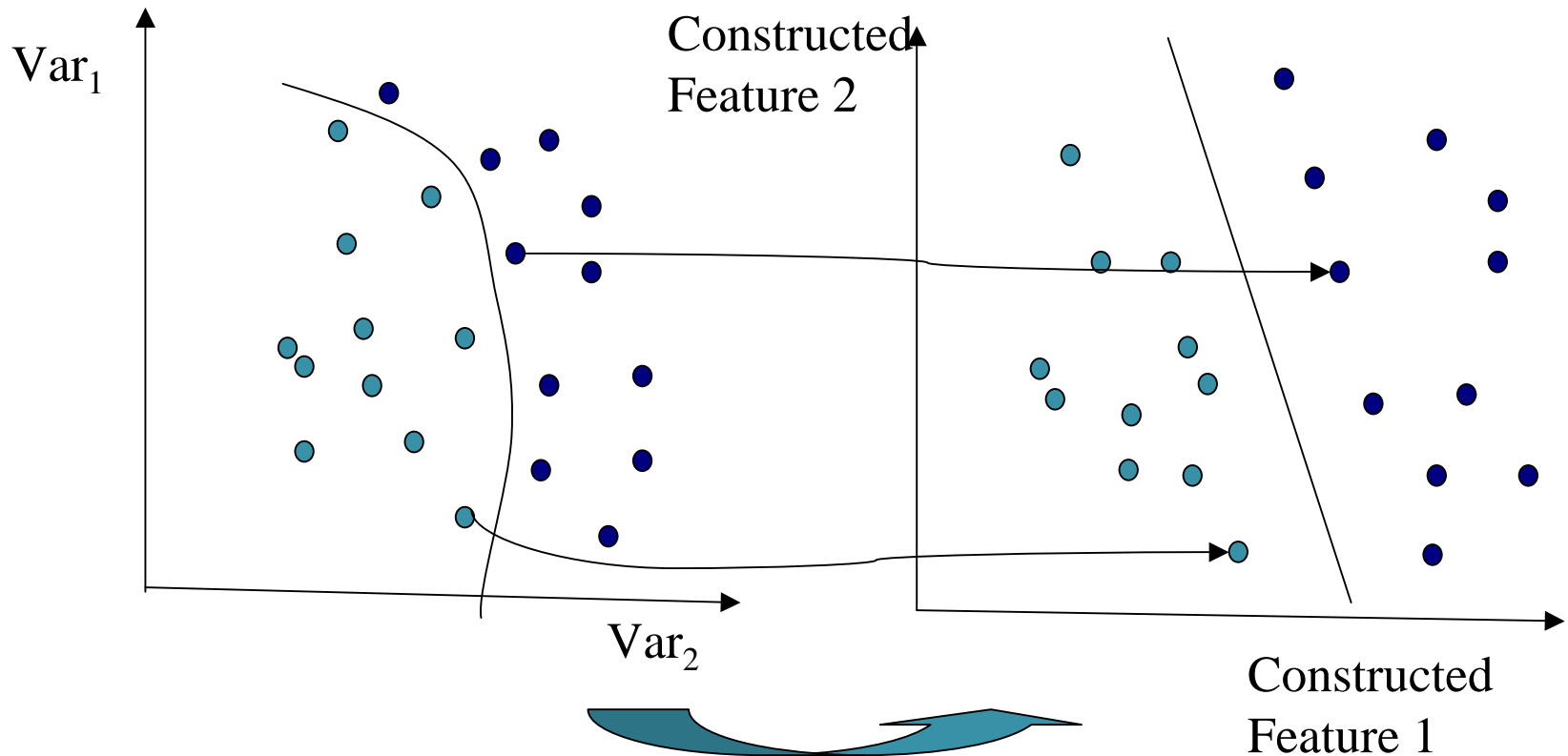
Labeled
Web Pages



Linear Support Vector Machine



Non-linear Support Vector Machine



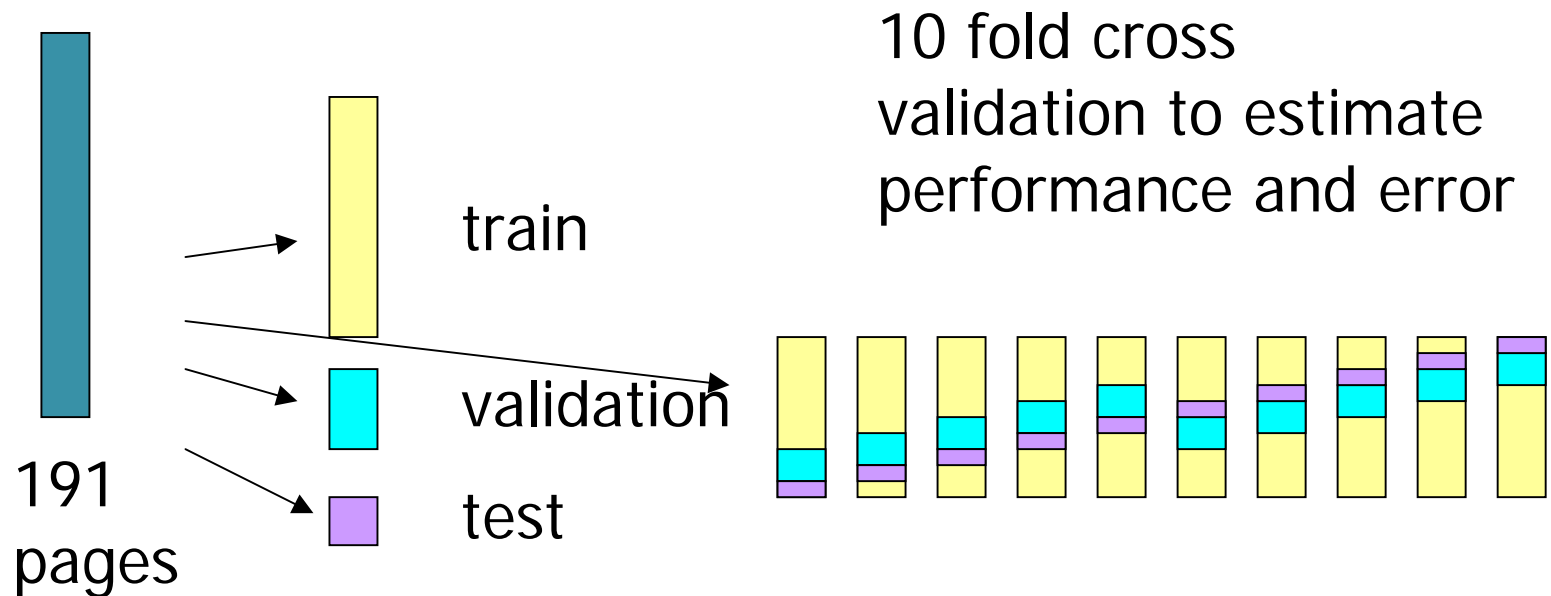
Find function $\Phi(x)$ to map to a different space

Quackometer.net

- Pioneering work but heuristic, unvalidated, and non-peer-reviewed quack detection tool.
- <http://www.quackometer.net>
- Looks for words in 5 dictionaries
 - Altmed terms – homeopathic, herbal.
 - Pseudoscientific words – toxins, superfoods.
 - Domain specific words – energy, vibration.
 - Skeptical words – placebo, flawed
 - Commerce terms – products, shipping
- Algorithm counts frequency of terms and applies a frequency threshold, then generates a score between 0 and 10.
- AUC for each 10 fold split.

Performance/Error Estimation

- Build a model.
- Estimate the performance of the methodology.



Google PageRank

- Compared PageRank within each topic to avoid bias in ranking situations.
- Observed Google rankings for top 30 results when query was applied and calculated an area under the curve.

Example - Google PageRank

- 1
- 2 quack page
- 3
- 4
- 5 quack page
- 6
- 7
- 8
- 9 quack page
- 10 quack page

Gold standard

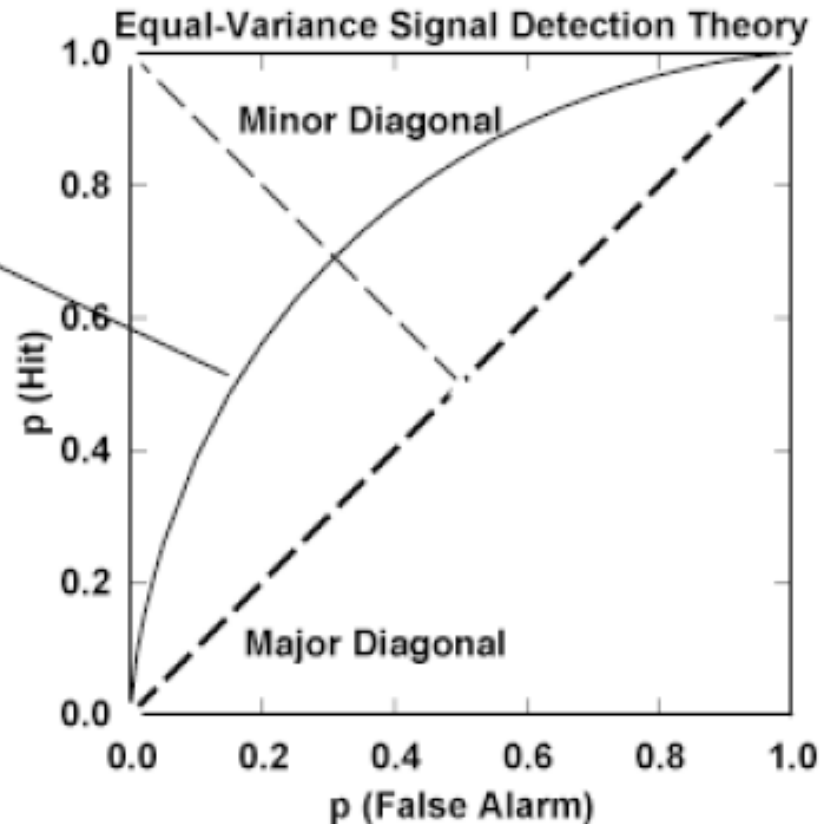
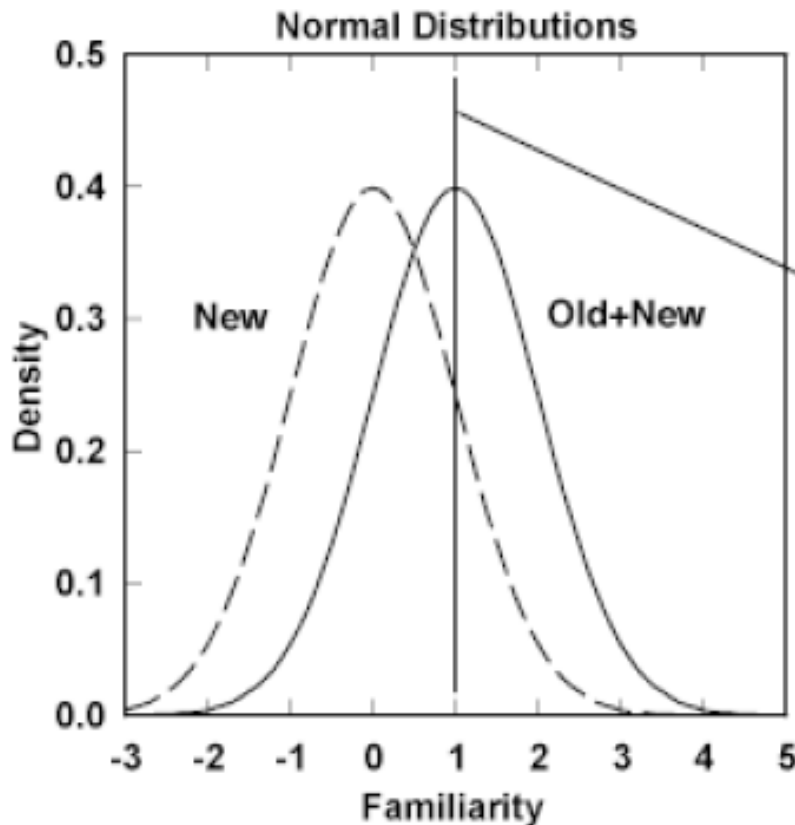
- 4 false claim pages

Area Under the Curve
– 0.71

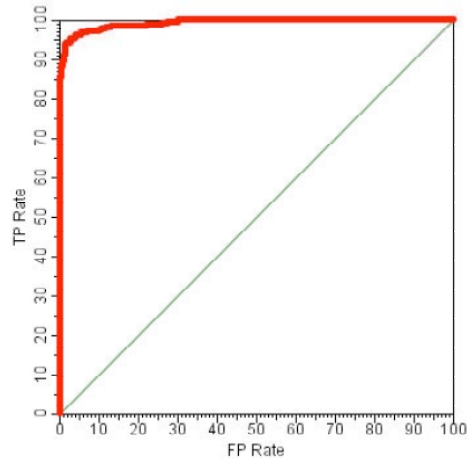
Interpreting Models

- Receiver Operating Curves.
- Area Under the Curve.

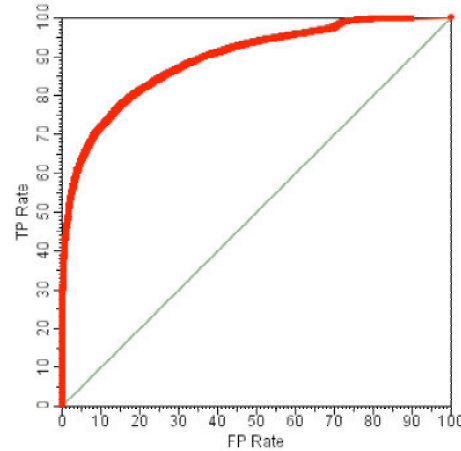
Centor RM. The Use of ROC Curves and Their Analyses. Med Decis Making 1991;11(2):102-106.



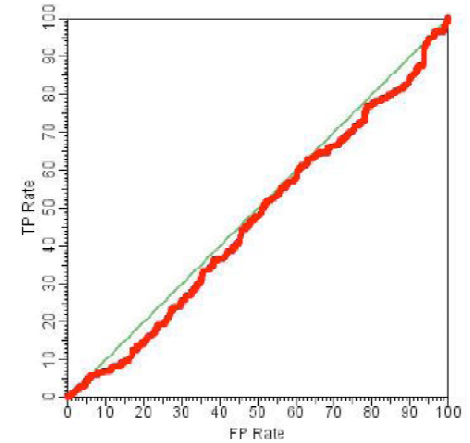
Example Receiver Operating Curves



Perfect Separation
Area Under the Curve
~ 1.0



Good Separation
Area Under the Curve ~
0.8



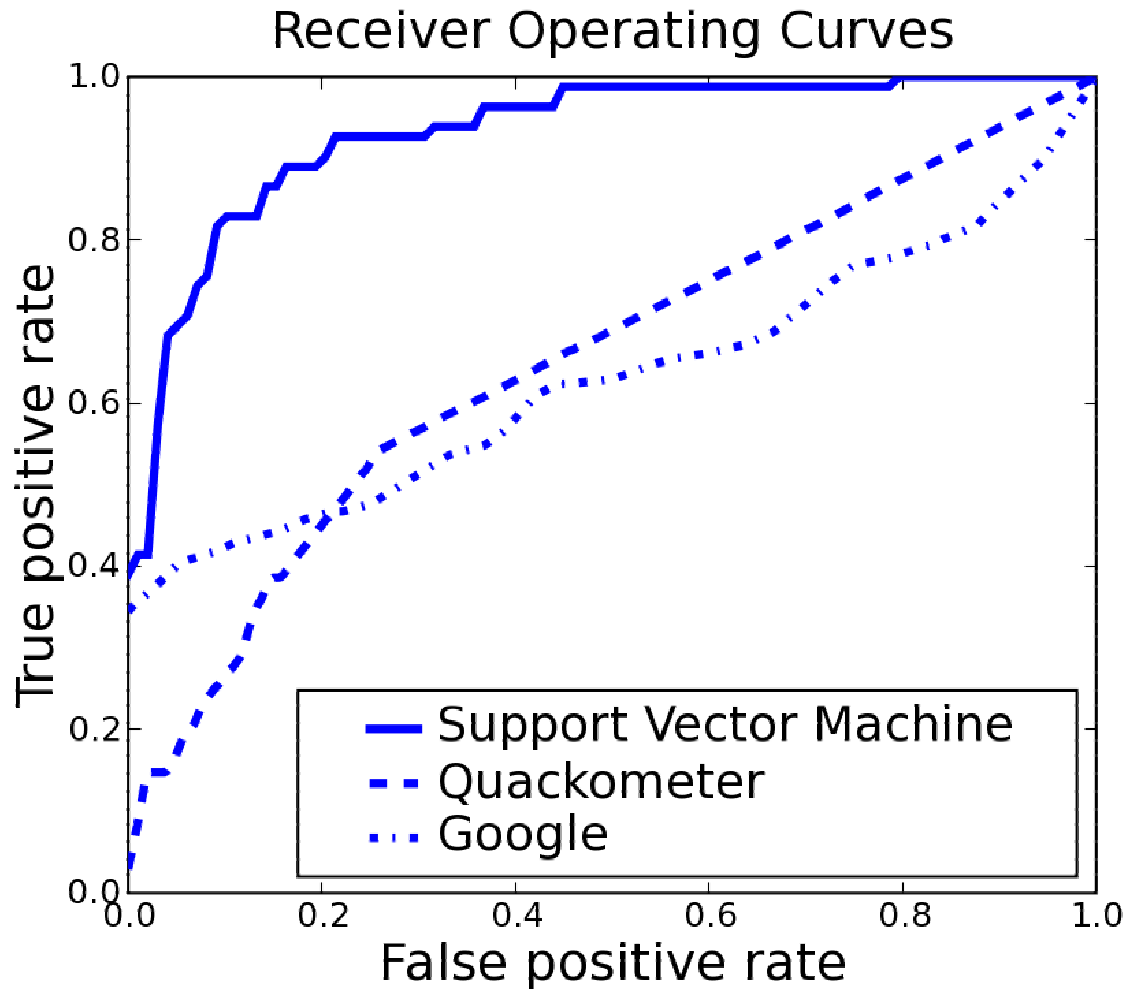
Uninformative (i.e.,
Random) Separation
Area Under the Curve
~ 0.5

Results

Model	Area Under the Curve
Machine Learning Models	0.93 (std. dev. 0.05)
Quackometer	0.67 (std. dev. 0.10)
Google*	0.63 (std. dev. 0.17)

* - Area under the curve calculated from Google ranks within each topic.

Receiver Operating Curves



Failure Analysis

- Support Vector Machine Rank
 - Success - Low quality page in top 10
 - Failure – Low quality page in bottom 20.
- Quackometer Score
 - Success – Low quality pages score 5 – 10.
 - Failure – Low quality pages score 0 – 5.
- Google
 - Success – Low quality page in bottom 20.
 - Failure – Low quality page in top 10.*
- Thresholds are chosen arbitrarily and for representation purposes only.

* - Consumers do not typically review results beyond the first page.

Failure Analysis Examples

Excerpts	Support Vector Rank	Quackometer Score	Google Rank
I am convinced that our mind and emotions are the deciding factor in the cure of cancer.	1 (S)	1 (F)	16 (S)
The hundreds of clinical studies conducted by many competent physicians around the world, including those directed by Dr. Ernesto Contreras Rodriguez at the Oasis of Hope Hospital hospital in , give us complete confidence that there is no danger.	3 (S)	0 (F)	9 (F)
The cure shows results almost immediately and lasts three weeks only. It is cheap and affordable for everybody and proved with 138 case studies.	3 (S)	8 (S)	3 (F)

Failure Analysis Examples

Excerpts	Support Vector Rank	Quackometer Score	Google Rank
Many advanced cancer patients are petrified of their tumor. This knee-jerk reaction is caused by orthodox medicine's focus on the highly profitable (and generally worthless) process of shrinking tumors.	1 (S)	1 (F)	18 (S)
IPT (Insulin Potentation Therapy) has an outstanding 135 doctor-year track record (115 years for cancer) over 72 years, and is ready for clinical trials and widespread use.	1 (S)	0 (F)	1 (F)
We are proud of these findings, which confirm that cellular medicine offers solutions for the most critical process in cancer development, the invasion of cancer cells to other organs in the body. Conventional medicine is powerless in this.	2 (S)	1 (F)	8 (F)

Advantages

- No need to develop explicit rating criteria.
- Allow automated application to any web page.



Key Note: State-of-the-art search engines cannot help health consumers identify fraudulent/dangerous pages

- Marginal relation between PageRank and quality.
- Gaming of PageRank, non-wisdom of the crowds.
- Search engines rank within one specific topic. The highest ranking document may have a low PageRank in that topic.
- Need machine learning filters on top of results to warn consumers of poor quality issues.

Price SL, Hersh WR. 1999.
Fallis D, Fricke M. 2002.
Fricke M, et al. 2005.
Griffiths KM, et al. 2005.
Tang TT, et al. 2006.

Limitations

- 8 unproven treatments only.
- Compared with Google only.
- Labeling limited to the two authors.
- Assume users would append “cancer” and “treatment” to their query.

Future Work

- Independent prospective validation.
- Expand space of unproven treatments and diseases.
- Expand labeling standard (reviewers, criteria, etc.)
- Identify terms that contribute to the labeling.
- Study feasibility of “gaming” the models.
- Build a practical system for consumer use.

Conclusions

- First validated study showing feasibility of pattern recognition filters that identify web pages that make unproven claims on the web.
- This technology is a first step toward automated mechanisms to protect patients from unproven/dangerous information on the web.

Acknowledgements

- Vanderbilt MD/ PhD Department
- Department of Biomedical Informatics
- NLM Training Grant