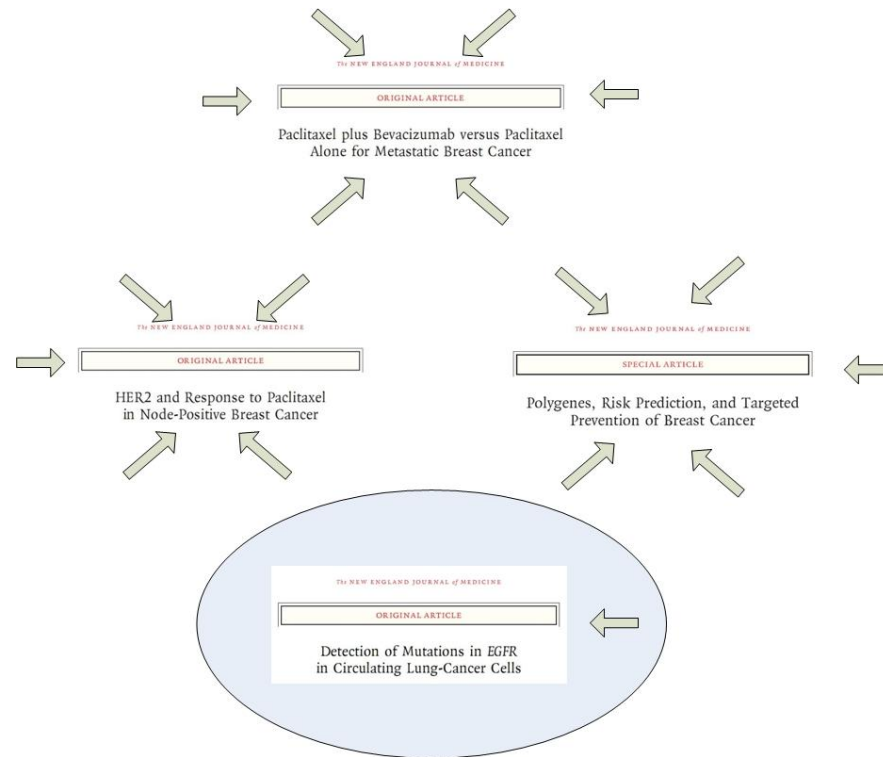

Improving Biomedical Information Retrieval Citation Metrics Using Machine Learning

**Lawrence Fu, PhD
January 26, 2009**

Introduction

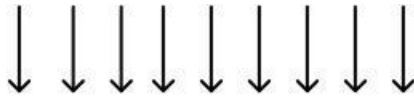
- MEDLINE: one of the most important informatics systems ever constructed
- Reflects importance of literature
- Complexity, size of literature make it difficult to find the most relevant, useful articles
- Automated, semi-automated tools have been developed to identify high quality articles
- Purpose: Improve the usability and performance of information retrieval techniques with machine learning methods

Focus 1: Topic-sensitivity



1. Analyze **topic-sensitivity** of evaluation methods for journals, articles, websites

Focus 2: Citation Count Prediction



The clinical significance of cerebrospinal fluid levels of kynurenine pathway metabolites and lactate in severe malaria.

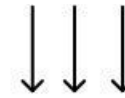
Medana IM, Hien TT, Day NP, Phu NH, Mai NT, Chu'ong LV, Chau TT, Taylor A, Salahifar H, Stocker K, Smythe G, Turner GD, Farrar J, White NJ, Hunt NH

Nuffield Department of Clinical Laboratory Sciences, Oxford-Wellcome Centre for Tropical and Infectious Diseases, John Radcliffe Hospital, Oxford OX3 9DU, United Kingdom. isabelle.medana@ndcs.ox.ac.uk

A retrospective study of 261 Vietnamese adults with severe malaria was conducted to determine the relationship between cerebrospinal fluid (CSF) levels of metabolites of the kynurenine pathway, the incidence of neurologic complications, and the disease outcome. Three metabolites were measured: the excitatory quinolinic acid (QA), the protective receptor antagonist kynurenic acid (KA), and the proinflammatory mediator picolinic acid (PA). These measurements were related prospectively to CSF lactate levels. QA and PA levels were elevated, compared with those of controls. There was no difference in the levels of KA between these groups. Although >40% of malaria patients had QA CSF concentrations in the micromolar range, there was no association with convulsions or depth of coma. Levels of QA and PA were associated significantly with death, but a multivariate analysis suggested that these elevations were a consequence of impaired renal function. CSF lactate remained an independent and significant predictor of poor outcome.

PMID: 1188422 [PubMed - indexed for MEDLINE]

Citation Count = 1000



The clinical significance of cerebrospinal fluid levels of kynurenine pathway metabolites and lactate in severe malaria.

Medana IM, Hien TT, Day NP, Phu NH, Mai NT, Chu'ong LV, Chau TT, Taylor A, Salahifar H, Stocker K, Smythe G, Turner GD, Farrar J, White NJ, Hunt NH

Nuffield Department of Clinical Laboratory Sciences, Oxford-Wellcome Centre for Tropical and Infectious Diseases, John Radcliffe Hospital, Oxford OX3 9DU, United Kingdom. isabelle.medana@ndcs.ox.ac.uk

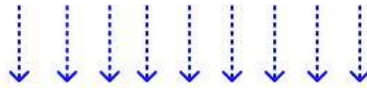
A retrospective study of 261 Vietnamese adults with severe malaria was conducted to determine the relationship between cerebrospinal fluid (CSF) levels of metabolites of the kynurenine pathway, the incidence of neurologic complications, and the disease outcome. Three metabolites were measured: the excitatory quinolinic acid (QA), the protective receptor antagonist kynurenic acid (KA), and the proinflammatory mediator picolinic acid (PA). These measurements were related prospectively to CSF lactate levels. QA and PA levels were elevated, compared with those of controls. There was no difference in the levels of KA between these groups. Although >40% of malaria patients had QA CSF concentrations in the micromolar range, there was no association with convulsions or depth of coma. Levels of QA and PA were associated significantly with death, but a multivariate analysis suggested that these elevations were a consequence of impaired renal function. CSF lactate remained an independent and significant predictor of poor outcome.

PMID: 1188422 [PubMed - indexed for MEDLINE]

Citation Count = 3

2. Is it feasible to predict the citation count of an article using only information available at the time of publication?

Focus 3: Automatic Classification of Citations



The clinical significance of cerebrospinal fluid levels of kynurenine pathway metabolites and lactate in severe malaria.

Medana IM, Hien TT, Day NP, Phu NH, Mai NT, Chu'ong LV, Chau TT, Taylor A, Salahifar H, Stocker R, Smythe G, Turner GD, Farrar J, White NJ, Hunt RH.

Nuffield Department of Clinical Laboratory Sciences, Oxford-Wellcome Centre for Tropical and Infectious Diseases, John Radcliffe Hospital, Oxford OX3 9DU, United Kingdom. isabelle.medana@ndcls.ox.ac.uk

A retrospective study of 261 Vietnamese adults with severe malaria was conducted to determine the relationship between cerebrospinal fluid (CSF) levels of metabolites of the kynurenine pathway, the incidence of neurologic complications, and the disease outcome. These metabolites were measured: the excitotoxin quinolinic acid (QA); the protective receptor antagonist kynurenic acid (KA); and the proinflammatory mediator picolinic acid (PA). These measurements were related prospectively to CSF lactate levels. QA and PA levels were elevated, compared with those of controls. There was no difference in the levels of KA between these groups. Although >40% of malaria patients had QA CSF concentrations in the micromolar range, there was no association with convulsions or depth of coma. Levels of QA and PA were associated significantly with death, but a multivariate analysis suggested that these elevations were a consequence of impaired renal function. CSF lactate remained an independent and significant predictor of poor outcome.

PMID: 11885422 [PubMed - indexed for MEDLINE]

Instrumental
Citation Count = 0



The clinical significance of cerebrospinal fluid levels of kynurenine pathway metabolites and lactate in severe malaria.

Medana IM, Hien TT, Day NP, Phu NH, Mai NT, Chu'ong LV, Chau TT, Taylor A, Salahifar H, Stocker R, Smythe G, Turner GD, Farrar J, White NJ, Hunt RH.

Nuffield Department of Clinical Laboratory Sciences, Oxford-Wellcome Centre for Tropical and Infectious Diseases, John Radcliffe Hospital, Oxford OX3 9DU, United Kingdom. isabelle.medana@ndcls.ox.ac.uk

A retrospective study of 261 Vietnamese adults with severe malaria was conducted to determine the relationship between cerebrospinal fluid (CSF) levels of metabolites of the kynurenine pathway, the incidence of neurologic complications, and the disease outcome. These metabolites were measured: the excitotoxin quinolinic acid (QA); the protective receptor antagonist kynurenic acid (KA); and the proinflammatory mediator picolinic acid (PA). These measurements were related prospectively to CSF lactate levels. QA and PA levels were elevated, compared with those of controls. There was no difference in the levels of KA between these groups. Although >40% of malaria patients had QA CSF concentrations in the micromolar range, there was no association with convulsions or depth of coma. Levels of QA and PA were associated significantly with death, but a multivariate analysis suggested that these elevations were a consequence of impaired renal function. CSF lactate remained an independent and significant predictor of poor outcome.

PMID: 11885422 [PubMed - indexed for MEDLINE]

Instrumental
Citation Count = 3

3. Is it feasible to automatically identify instrumental citations?

Focus 1: Topic Sensitivity

- Purpose: analyze topic-sensitivity of evaluation methods
- Previous work focused on overall performance
- Performance on specific topics unknown
- Benefits:
 - Raise awareness of issue
 - Provide alternatives that consider topic or not vulnerable to this issue

Evaluating Journal Quality

■ Impact Factor

- Measures citation rate regardless of publication size or frequency

$$= \frac{\text{Number of citations in year } y \text{ to journal items published in years } (y - 1) \text{ and } (y - 2)}{\text{Number of journal articles published in years } (y - 1) \text{ and } (y - 2)}$$

■ Topic-Specific Impact Factor

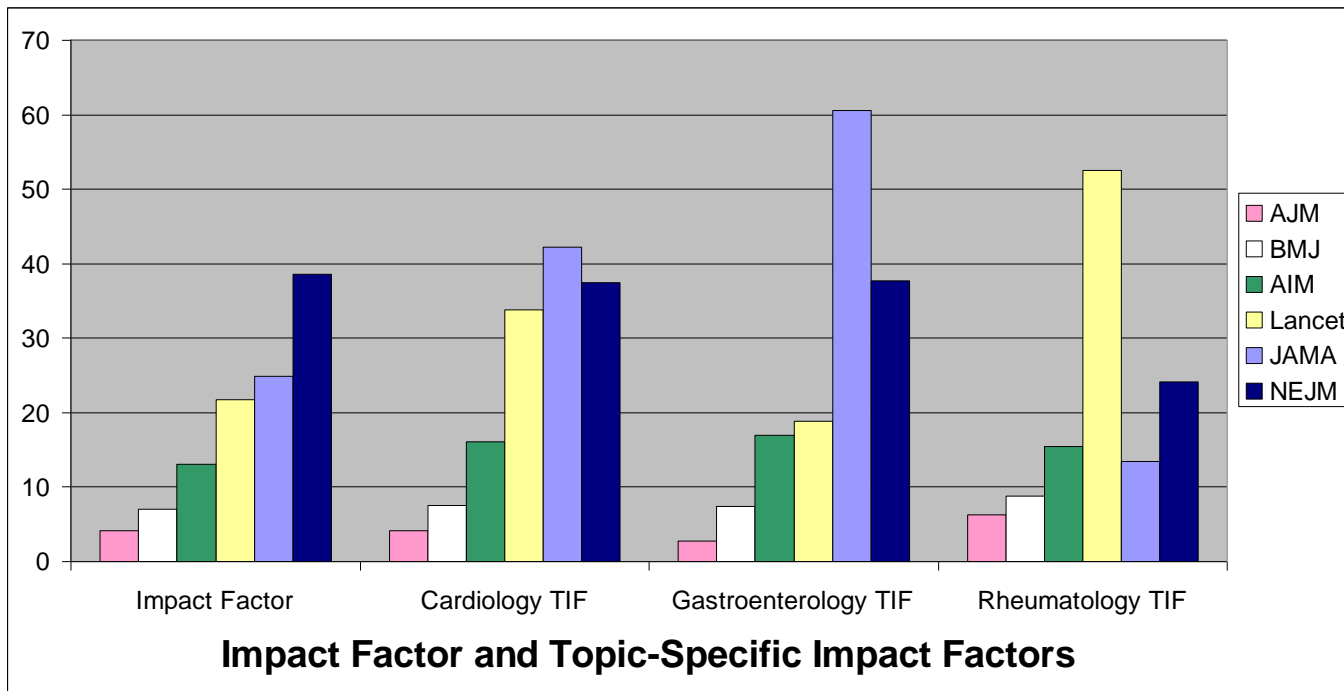
- Restrict to articles related to given topic
- “Impact factor” of subset of journal’s articles

Experimental Design

- Identified matching MEDLINE records
 - Topic: 8 general topics
 - Journal: 6 journals
 - Year: 2004, 2003
- Retrieved citation counts and journal impact factor from ISI Web of Science
- Calculated topic-specific impact factors for all journals, topics, and years

Results

- Variability shown by ranking reversals
 - Cases where higher impact journal had lower topic-specific impact factor
 - 10 reversals out of 120 comparison



Results

- Wide variability shown by differences between impact factor and topic-specific impact factors

Topic	Minimum	Median	Maximum	Interquartile Range
Cardiology	0.09	2.04	17.35	11.58
Endocrinology	0.56	2.09	25.99	15
Gastroenterology	0.33	2.15	35.72	2.92
Hematology	1.31	5.02	10.96	7.53
Medical Oncology	0.23	1.46	10.75	5.61
Nephrology	0.13	6.04	10.64	5.55
Pulmonary Disease	0.45	0.99	11.64	5.1
Rheumatology	1.73	6.86	30.79	12.38

Interquartile range is a measure of dispersion and is the difference between the first and third quartiles.

Finding High-quality Articles

■ Clinical Query Filters

- Manually constructed Boolean queries of terms from MEDLINE record
- Optimized for sensitivity, specificity
- (randomized controlled trial [Publication Type] OR (randomized [Title/Abstract] AND controlled [Title/Abstract] AND trial [Title/Abstract]))

■ Machine Learning Method

- Support Vector Machine (SVM) models
- Performs well in categorizing text and identifying high-quality articles

Experimental Design

- ACP Journal Club: corpus, gold standard
 - Experts review the best journals in internal medicine, identify high-quality articles
- Selected 18 topics based on MeSH terms
- Compute topic-specific performance
- Compare to overall performance

Clinical Query Filters Results

- Performance measured by sensitivity, specificity
- Differences between overall and topic-specific sensitivity, specificity varied greatly

Optimized for	Category	Sensitivity				Specificity			
		Min	Median	Max	IQR	Min	Median	Max	IQR
Sensitivity	Diagnosis	0.02	0.02	0.15	0.0013	0.015	0.087	0.23	0.097
	Etiology	0.028	0.07	0.07	0	0.00047	0.059	0.22	0.10
	Prognosis	0.031	0.1	0.57	0.15	0.0029	0.053	0.18	0.042
	Treatment	0.0035	0.01	0.026	0.0025	0.0027	0.030	0.17	0.053
Specificity	Diagnosis	-	-	-	-	-	-	-	-
	Etiology	0.16	0.34	0.49	0.28	0.0066	0.13	0.31	0.086
	Prognosis	0.11	0.24	0.52	0.33	0.030	0.099	0.22	0.035
	Treatment	0.034	0.053	0.07	0.023	0.00037	0.048	0.13	0.033

SVM Models Results

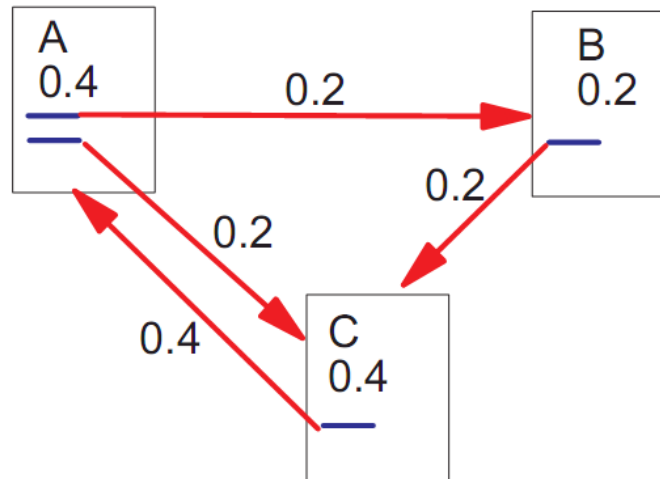
- Performance measured by area under the receiver operating characteristics curve (AUC)
- More stable results for differences between overall and topic-specific AUC

Category	Min.	Median	Max.	IQR
Diagnosis	0.0083	0.038	0.04	0.012
Etiology	0.0027	0.028	0.13	0.05
Prognosis	0.0041	0.045	0.10	0.065
Treatment	0.00054	0.0040	0.041	0.0078

Evaluating Websites

■ PageRank

- High quality pages link to other high quality pages
- Models user behavior as random surfer that follows link arbitrarily or jumps randomly to another page



Experimental Design

- Selected sites and topics from WebBase
 - CDC: Genomics, NCBDDD, NCIDOD, NIP, Tobacco
 - NCI: Breast, Cervix, Colon, Lung, Prostate
- Computed PageRanks before, after topic isolation
- Measured similarity in rankings with Ksim metric
 - Fraction of consistent pairwise ranking comparisons between two sets of rankings

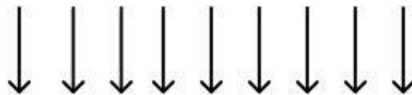
Results

- Stability of rankings dependent on how often pages linked to pages outside of topic

Domain	Topic	Ksim for Topic subset	Fraction of links within same topic
CDC	Genomics	0.97	0.85
	NCBDDD	0.87	0.71
	NCIDOD	0.79	0.76
	NIP	0.87	0.83
	Tobacco	0.94	0.87
NCI	Breast	0.71	0.32
	Cervix	0.74	0.42
	Colon	0.72	0.37
	Lung	0.76	0.36
	Prostate	0.7	0.32

Focus 2: Prediction of Citation Counts

- Citation count: higher quality papers receive more citations
- Simple, efficient, intuitive method
- Limitations: difficulties comparing papers for different topics or time periods



The clinical significance of cerebrospinal fluid levels of kynurenine pathway metabolites and lactate in severe malaria.

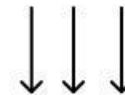
Medana JM, Hien TT, Day NP, Phu NH, Mai NT, Chu'ong LV, Chau TT, Taylor A, Salihifar H, Stocker R, Smythe G, Turner GD, Farrar J, White NJ, Hunt RH

Nuffield Department of Clinical Laboratory Sciences, Oxford-Wellcome Centre for Tropical and Infectious Diseases, John Radcliffe Hospital, Oxford OX3 9DU, United Kingdom. isabelle.medana@ndls.ox.ac.uk

A retrospective study of 261 Vietnamese adults with severe malaria was conducted to determine the relationship between cerebrospinal fluid (CSF) levels of metabolites of the kynurenine pathway, the incidence of neurologic complications, and the disease outcome. Three metabolites were measured: the excitotoxin quinolinic acid (QA); the protective receptor antagonist kynurenic acid (KA); and the proinflammatory mediator picolinic acid (PA). These measurements were related prospectively to CSF lactate levels. QA and PA levels were elevated, compared with those of controls. There was no difference in the levels of KA between these groups. Although >40% of malaria patients had QA CSF concentrations in the micromolar range, there was no association with convulsions or depth of coma. Levels of QA and PA were associated significantly with death, but a multivariate analysis suggested that these elevations were a consequence of impaired renal function. CSF lactate remained an independent and significant predictor of poor outcome.

PMID: 11885422 (PubMed - release for MEDLINE)

Citation Count = 1000



The clinical significance of cerebrospinal fluid levels of kynurenine pathway metabolites and lactate in severe malaria.

Medana JM, Hien TT, Day NP, Phu NH, Mai NT, Chu'ong LV, Chau TT, Taylor A, Salihifar H, Stocker R, Smythe G, Turner GD, Farrar J, White NJ, Hunt RH

Nuffield Department of Clinical Laboratory Sciences, Oxford-Wellcome Centre for Tropical and Infectious Diseases, John Radcliffe Hospital, Oxford OX3 9DU, United Kingdom. isabelle.medana@ndls.ox.ac.uk

A retrospective study of 261 Vietnamese adults with severe malaria was conducted to determine the relationship between cerebrospinal fluid (CSF) levels of metabolites of the kynurenine pathway, the incidence of neurologic complications, and the disease outcome. Three metabolites were measured: the excitotoxin quinolinic acid (QA); the protective receptor antagonist kynurenic acid (KA); and the proinflammatory mediator picolinic acid (PA). These measurements were related prospectively to CSF lactate levels. QA and PA levels were elevated, compared with those of controls. There was no difference in the levels of KA between these groups. Although >40% of malaria patients had QA CSF concentrations in the micromolar range, there was no association with convulsions or depth of coma. Levels of QA and PA were associated significantly with death, but a multivariate analysis suggested that these elevations were a consequence of impaired renal function. CSF lactate remained an independent and significant predictor of poor outcome.

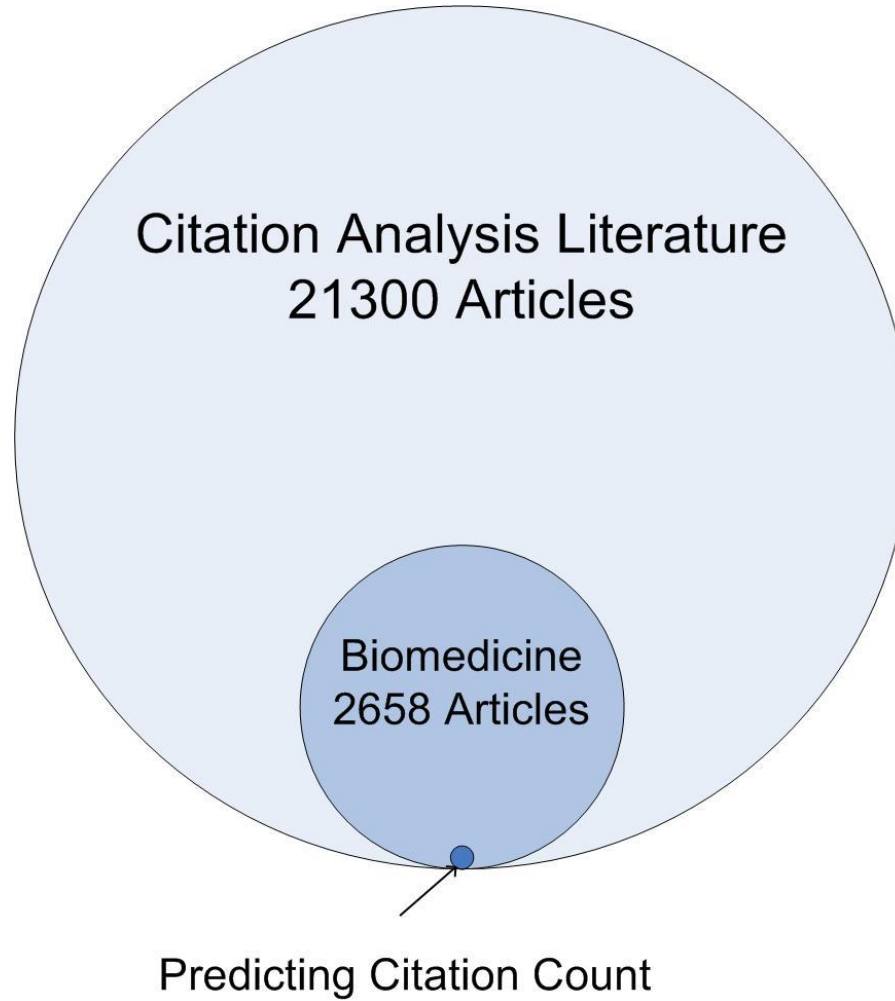
PMID: 11885422 (PubMed - release for MEDLINE)

Citation Count = 3

Purpose

- Can we predict citation count as a measure of the long term impact of papers at the time of publication?
- Benefits:
 - Accelerate research
 - Improve understanding of factors influencing citation behavior

Related Work



Framing the Problem: Text Categorization

The image displays three web interfaces used for text categorization. On the left is the ISI Web of Knowledge search interface, featuring a search bar and multiple fields for refining results. In the center is the PubMed interface, showing a search bar and navigation tabs. On the right is the Evidence-Based Medicine interface, which includes a sidebar with 'Subscriptions', 'CURRENT ISSUE', 'Powerpoints', and 'SPECIAL FEATURES'. A green box highlights a section of the PubMed interface titled 'The NIH Public Access Policy May Affect You'.

1. Build or utilize existing training corpora

Framing the Problem: Text Categorization

1: J Infect Dis. 2002 Mar 1;185(5):650-6. Epub 2002 Feb 14.

[Related Articles](#), [Links](#)

The University of
Chicago Press

The clinical significance of cerebrospinal fluid levels of kynurenine pathway metabolites and lactate in severe malaria.

Medana IM, Hien TT, Day NP, Phu NH, Mai NT, Chu'ong LV, Chau TT, Taylor A, Salahifar H, Stocker R, Smythe G, Turner GD, Farrar J, White NJ, Hunt NH

Nuffield Department of Clinical Laboratory Sciences, Oxford-Wellcome Centre for Tropical and Infectious Diseases

A retrospective study of 261 Vietnamese adults with severe malaria was conducted to determine the relationship between cerebrospinal fluid (CSF) levels of metabolites of the kynurenine pathway, the incidence of neurologic complications, and the disease outcome. Three metabolites were measured: the excitotoxin quinolinic acid (QA), the protective receptor antagonist kynurenic acid (KA), and the proinflammatory mediator picolinic acid (PA). These measurements were related prospectively to CSF lactate levels. QA and PA levels were elevated, compared with those of controls. There was no difference in the levels of KA between these groups. Although >40% of malaria patients had QA CSF concentrations in the micromolar range, there was no association with convulsions or depth of coma. Levels of QA and PA were associated significantly with death, but a multivariate analysis suggested that these elevations were a consequence of impaired renal function. CSF lactate remained an independent and significant predictor of poor outcome.

Publication Types:

- Clinical Trial
- Randomized Controlled Trial

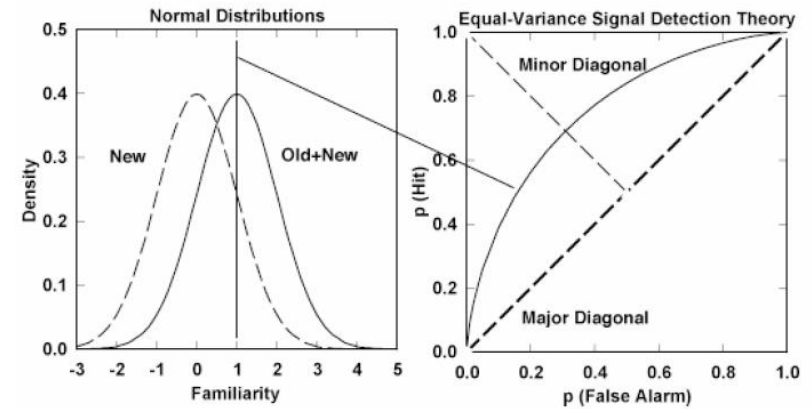
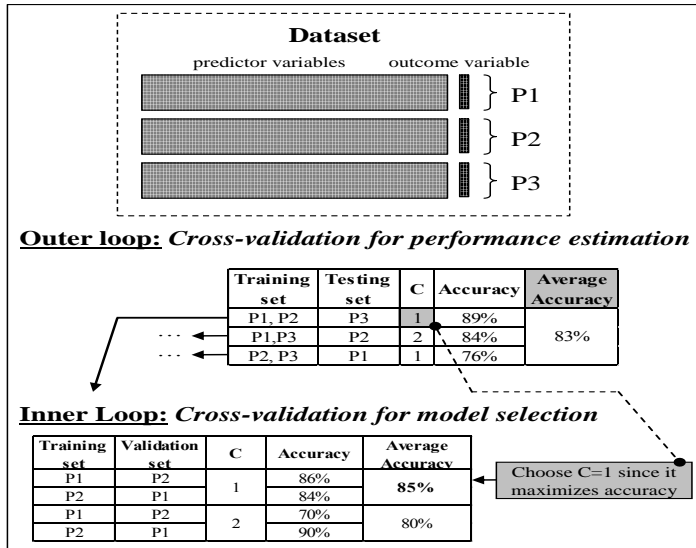
MeSH Terms:

- Malaria, Cerebral/cerebrospinal fluid*
- Malaria, Cerebral/drug therapy
- Malaria, Cerebral/parasitology

PMID: 11865422 [PubMed - indexed for MEDLINE]

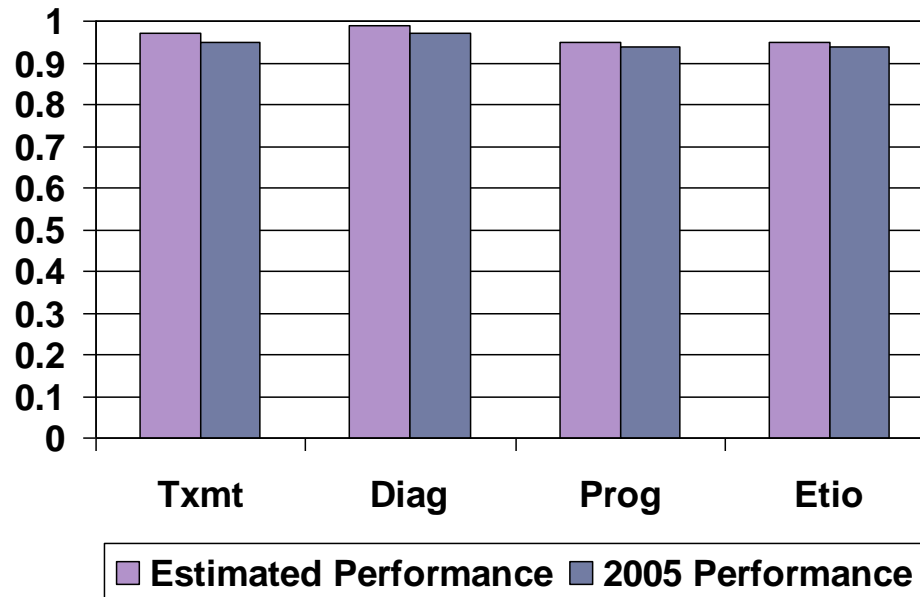
2. Simple document representations (typically stemmed and weighted words in title, abstract, MeSH terms)

Framing the Problem: Text Categorization



4. Evaluate models' performances with **nested cross-validation** and area under the receiver operating characteristics curve (**AUC**)

Framing the Problem: Text Categorization



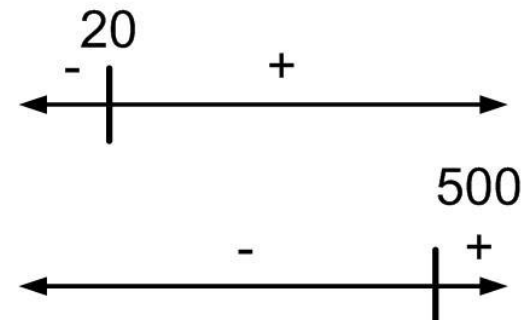
5. Evaluate performance **prospectively** & compare to prior cross-validation estimates

Predictive Features, Response Variable

Feature	Source	Representation	
Article Title Article Abstract MeSH terms	MEDLINE	~20000 features after processing	Content Features
Number of citations for first author Number of citations for last author Number of articles for first author Number of articles for last author Number of Authors Number of Institutions Publication Type Journal Impact Factor	Web of Science: not publically available, has to be extracted	1 integer value 1 binary value 1 continuous value	Bibliometric Features
Quality of First Author's Institution	www.arwu.org	4 values	Response Variable
Citation Count	Web of Science	1 integer value	

Convert to binary label based on citation thresholds

- > 20 : mildly influential
- > 50 : relatively influential
- > 100: influential
- > 500: extremely influential



Corpus Construction

“N Engl J Med”[Journal] AND (“1991”[PDAT] :
“1994”[PDAT]) AND (“Cardiology”[MeSH])

8 Topics from Internal Medicine
6 Journals
Published in 1991-1994

1. Query PubMed for articles according to selection criteria

Title
Abstract
MeSH terms

1: J Infect Dis. 2002 Mar 1;185(5):650-6. Epub 2002 Feb 14. [Related Articles, Links](#)

The University of Chicago Press

The clinical significance of cerebrospinal fluid levels of kynurenine pathway metabolites and lactate in severe malaria.

Medana IM, Hien TT, Day NP, Phu NH, Mai NT, Chu'ong LV, Chau TT, Taylor A, Salahifar H, Stocker R, Smythe G, Turner GD, Farrar J, White NJ, Hunt NL

Nuffield Department of Clinical Laboratory Sciences, Oxford-Wellcome Centre for Tropical and Infectious Diseases

A retrospective study of 261 Vietnamese adults with severe malaria was conducted to determine the relationship between cerebrospinal fluid (CSF) levels of metabolites of the kynurenine pathway, the incidence of neurologic complications, and the disease outcome. These metabolites were measured: the excitotoxin quinolinic acid (QA), the protective receptor antagonist kynurenic acid (KA), and the proinflammatory mediator picolinic acid (PA). These measurements were related prospectively to CSF lactate levels. QA and PA levels were elevated, compared with those of controls. There was no difference in the levels of KA between these groups. Although >40% of malaria patients had QA CSF concentrations in the nanomolar range, there was no association with consciousness or depth of coma. Levels of QA and PA were associated significantly with death, but a multivariate analysis suggested that these elevations were a consequence of impaired renal function. CSF lactate remained an independent and significant predictor of poor outcome.

Publication Types:
• Clinical Trial
• Randomized Controlled Trial

MeSH terms:
• Malaria, Cerebro[neuro]spinal fluid*
• Malaria, Cerebral/drug therapy
• Malaria, Cerebral/parasitology

PMID: 11865422 [PubMed - indexed for MEDLINE]

2. Download content features from MEDLINE using Python scripts

Corpus Construction

ISI Web of KnowledgeSM Take the next step

All Databases Select a Database Web of Science Additional Resources

Search Search History Marked List (0)

ALL DATABASES

<< Back to results list Record 1 of 2 >>

The clinical significance of cerebrospinal fluid levels of kynurenine pathway metabolites and lactate in severe malaria

find it NCBI Print Email Add to Marked List Save to EndNote@Web

Holdings Go Save to EndNote, RefMan, ProCite more options

Author(s): Medana IM, Nien TT, Day NP, Phu NH, Mai NTH, Chu'ong LV, Chau TTH, Taylor A, Salahifar H, Stocker R, Smythe G, Turner GDH, Farrar J, White NJ, Hunt NH

Source: JOURNAL OF INFECTIOUS DISEASES Volume: 185 Issue: 5 Pages: 650-656 Published: MAR 1 2002

Times Cited: 25 References: 28 Citation Map Beta

Results Author=(Medana IM) Timespan=All Years

Results: 22 Page 1 of 3 Go Sort by: Publication Date

Refine Results Search within results for

General Categories (Refine) SCIENCE & TECHNOLOGY (22) SOCIAL SCIENCES (1) more options / values

Subject Areas (Refine) NEUROSCIENCES & NEUROLOGY (20) PARASITOLOGY (19) IMMUNOLOGY (15) BIO-CHEMISTRY & MOLECULAR BIOLOGY (13) INFECTIOUS DISEASES (12) more options / values

Document Types Authors Source Titles Publication Years Languages

1. Title: Host vascular endothelial growth factor is trophic for Plasmodium falciparum-infected red blood cells
Author(s): Sachanonta, N, Medana, IM, Robert, S, et al.
Source: ASIAN PACIFIC JOURNAL OF ALLERGY AND IMMUNOLOGY Volume: 26 Issue: 1 Pages: 37-45 Published: 2008 Times Cited: 0 Find it

2. Title: Fatal cerebral malaria: distinct microvascular pathologies in children and adult patients
Author(s): Wassmer, SC, Medana, IM, Turner, GDH, et al.
Source: INTERNATIONAL JOURNAL FOR PARASITOLOGY Volume: 38 Pages: S44-S44 Published: 2008

Citing Articles

Title: The clinical significance of cerebrospinal fluid levels of kynurenine pathway metabolites and lactate in severe malaria
Author(s): Medana, IM
Source: JOURNAL OF INFECTIOUS DISEASES Volume: 185 Issue: 5 Pages: 650-656 Published: MAR 1 2002
Citation Map Beta

The above article has been cited by the articles listed below
Note: The Times Cited count is calculated across all Web of Science editions. More information.

Results: 25 Page 1 of 3 Go Sort by: Latest Date

Refine Results Search within results for

Subject Areas (Refine) PARASITOLOGY (2) IMMUNOLOGY (4) NEUROSCIENCES (4) INFECTIOUS DISEASES (3) BIO-CHEMISTRY & MOLECULAR BIOLOGY (2) more options / values

Document Types (Refine) ARTICLE (11) REVIEW (3) PROCEEDINGS PAPER (4) MEETING ABSTRACT (2) more options / values

Authors Source Titles Publication Years Institutions Languages Countries/Territories

1. Title: Tryptophan in physiology and pathophysiology
Author(s): Hunt NH-Source: FREE RADICAL RESEARCH Volume: 42 Pages: S25-S25 Supplement: Suppl. 1 Published: JUL 2008 Times Cited: 0 Find it

2. Title: In-hospital risk estimation in children with Malaria - Early predictors of morbidity and mortality
Author(s): Winler AS, Sainthofer G, Heibok R, et al-Source: JOURNAL OF TROPICAL PEDIATRICS Volume: 54 Issue: 3 Pages: 184-191 Published: JUN 2008 Times Cited: 0 Find it

3. Title: Immunosuppression routed via the kynurenine pathway: A biochemical and pathophysiologic approach
Author(s): Gonzalez A, Vero N, Alegre E, et al-Source: ADVANCES IN CLINICAL CHEMISTRY, VOL 45 Book Series: ADVANCES IN CLINICAL CHEMISTRY Volume: 45 Pages: 155-197 Published: 2008 Times Cited: 1 Find it

4. Title: Characterization of an indoleamine 2,3-dioxygenase-like protein found in humans and mice
Author(s): Ball HJ, Sanchez-Perez A, Weller S, et al-Source: GENE Volume: 396 Issue: 1 Pages: 203-213 Published: JUL 1 2007 Times Cited: 12 Find it

- 3. Download **bibliometric features** from Web of Science
- Simulate user session with Python scripts
- Manually extract features for difficult cases

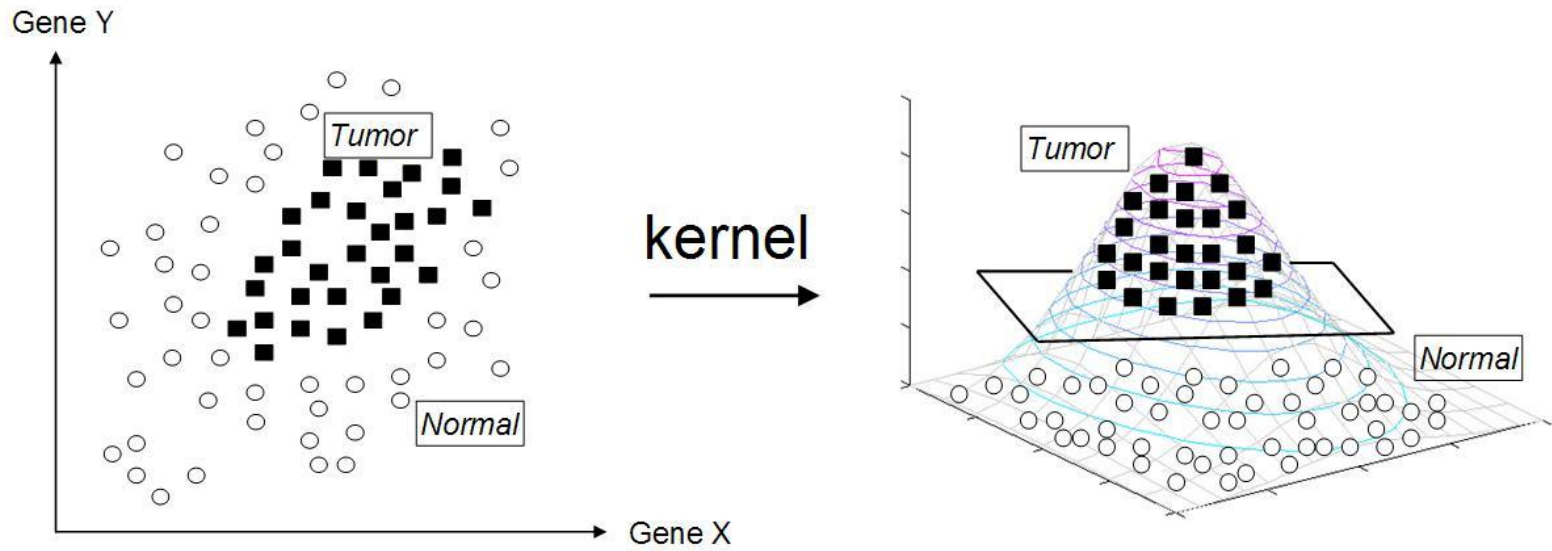
Document Representation

- Bag of words approach
- Removed PubMed stopwords (“a”, “the”, etc.)
- Porter stemming (activates, activating → activat)
- Weighting: log frequency with redundancy
- Bibliometric features scaled between 0 and 1
- Document represented as a vector of weights

$$\begin{bmatrix} .111 & .033 & \dots & 0 \\ .111 & 0 & \dots & .45 \\ \vdots & \vdots & \ddots & \vdots \\ .222 & .077 & \dots & 0 \end{bmatrix}$$

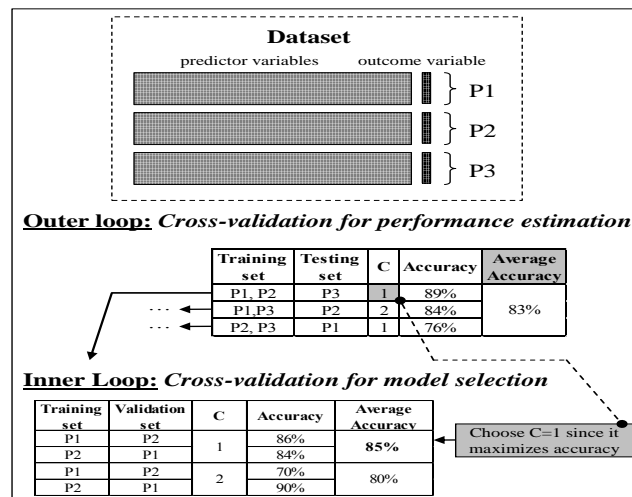
Learning Method

- Support Vector Machine (SVM) models
 - Kernel function maps input space to higher-dimensional feature space
 - Hyperplane calculated to separate classes of data
 - Performs well in categorizing text and identifying high-quality articles



Model Selection

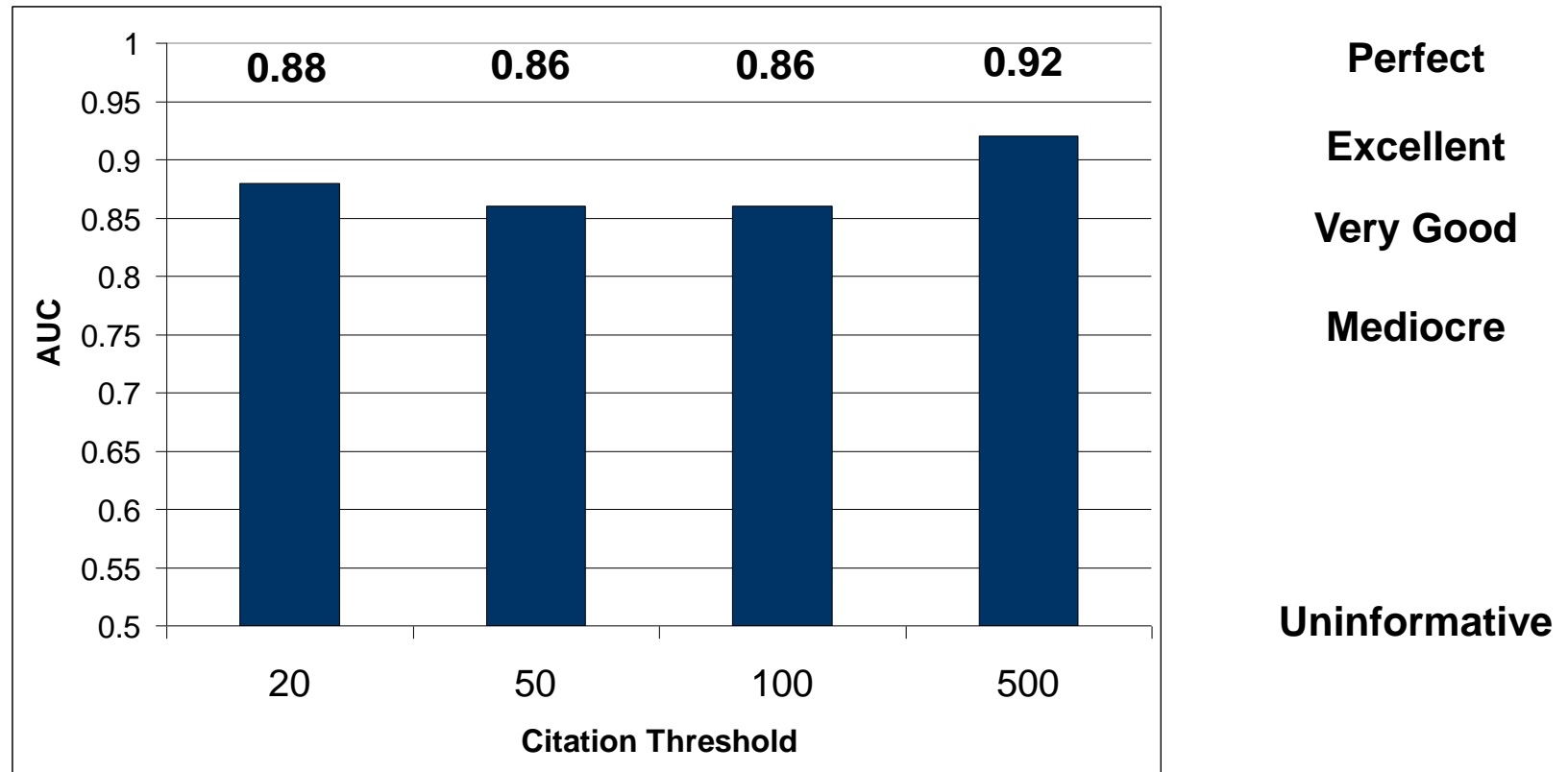
- 5-fold nested cross validation
- Optimized parameters
 - Cost: [.1, .2, .4, .7, .9, 1, 5, 10, 20]
 - Degree: [1, 2, 3, 4, 5, 8]
- Performance metric: Area under the receiver operating characteristic curve (AUC)



Results: Predictivity

- Possible to predict citation count with high predictivity with only information available at publication

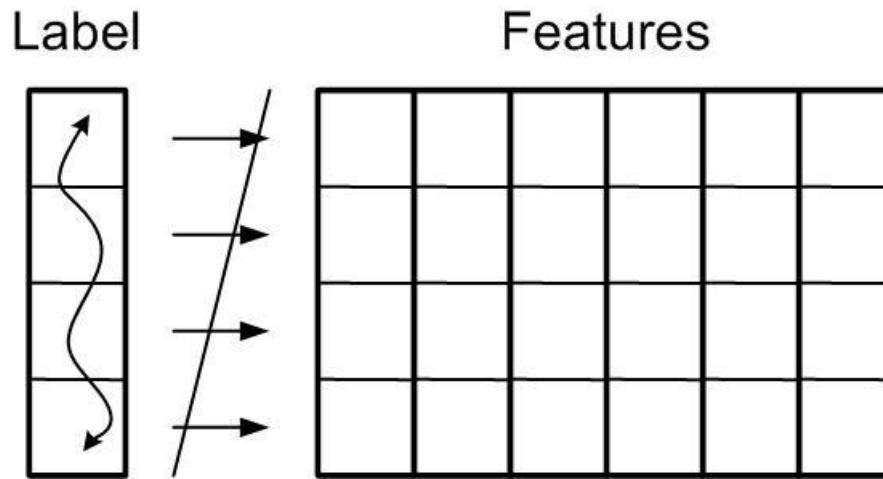
Classification Performance



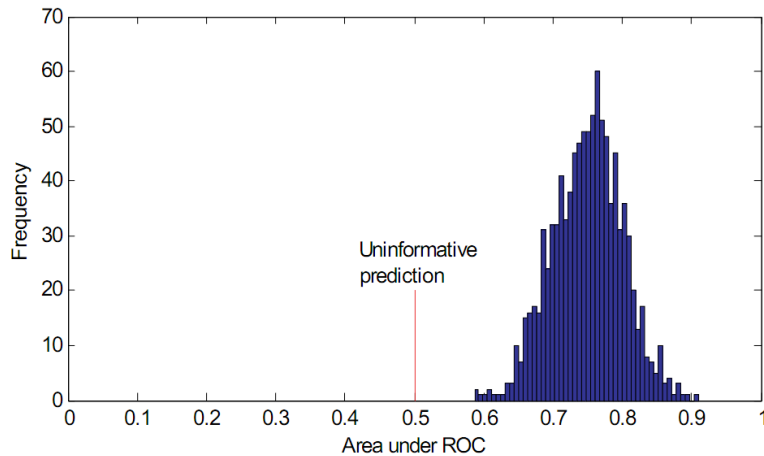
Results: Testing for Overfitting

■ Label reshuffling

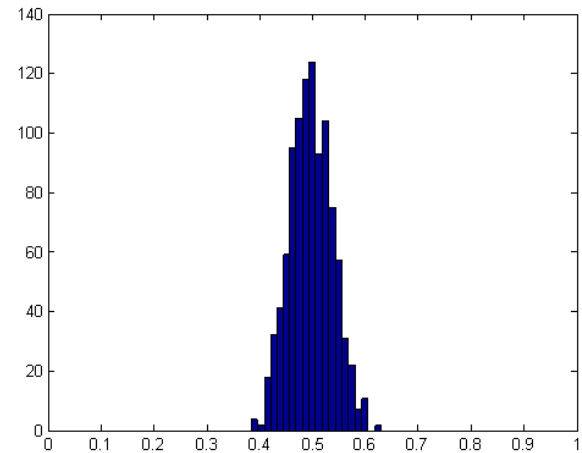
- Randomly reshuffle label to eliminate connection between features and label
- Retrain model, repeat multiple times
- AUC = 0.5 indicates no overfitting



Results: Testing for Overfitting



Overfitting

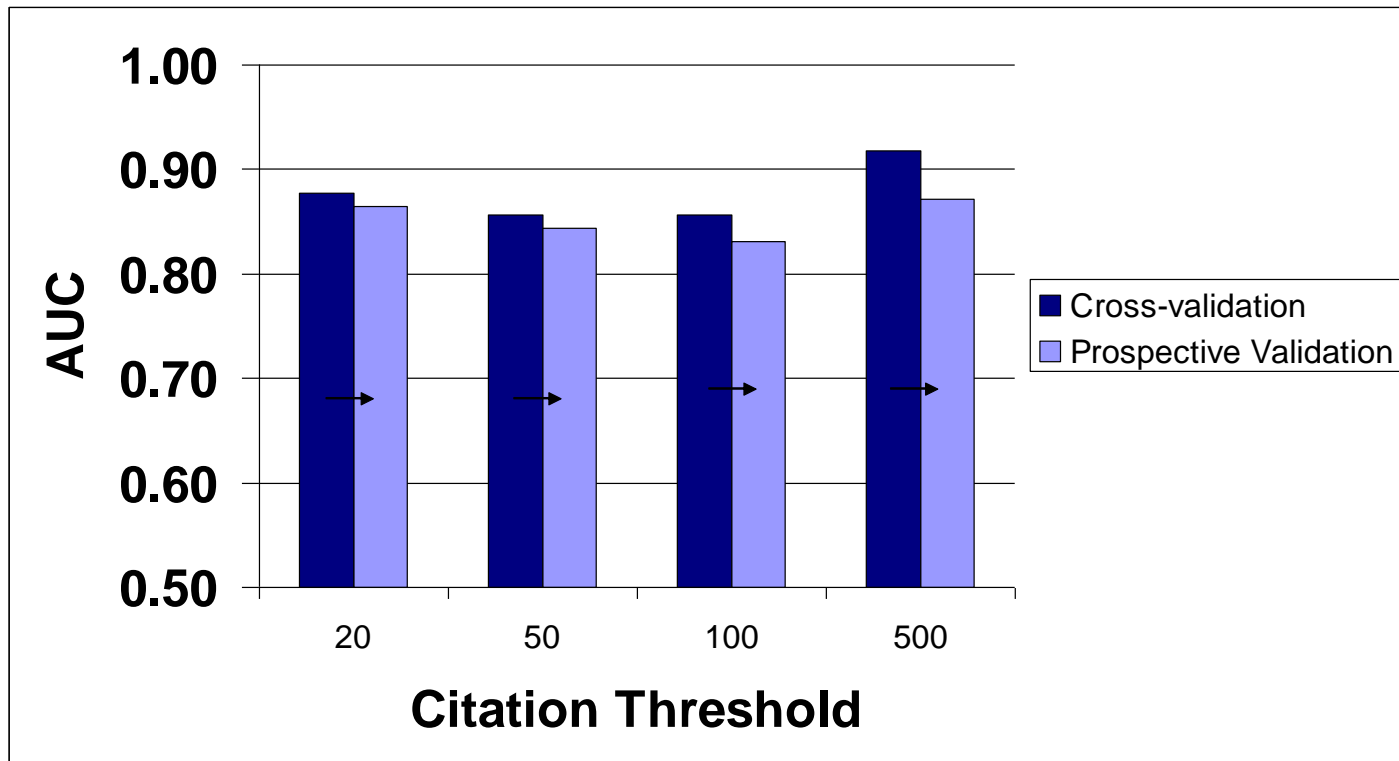


No Overfitting

Results: Testing for Overfitting

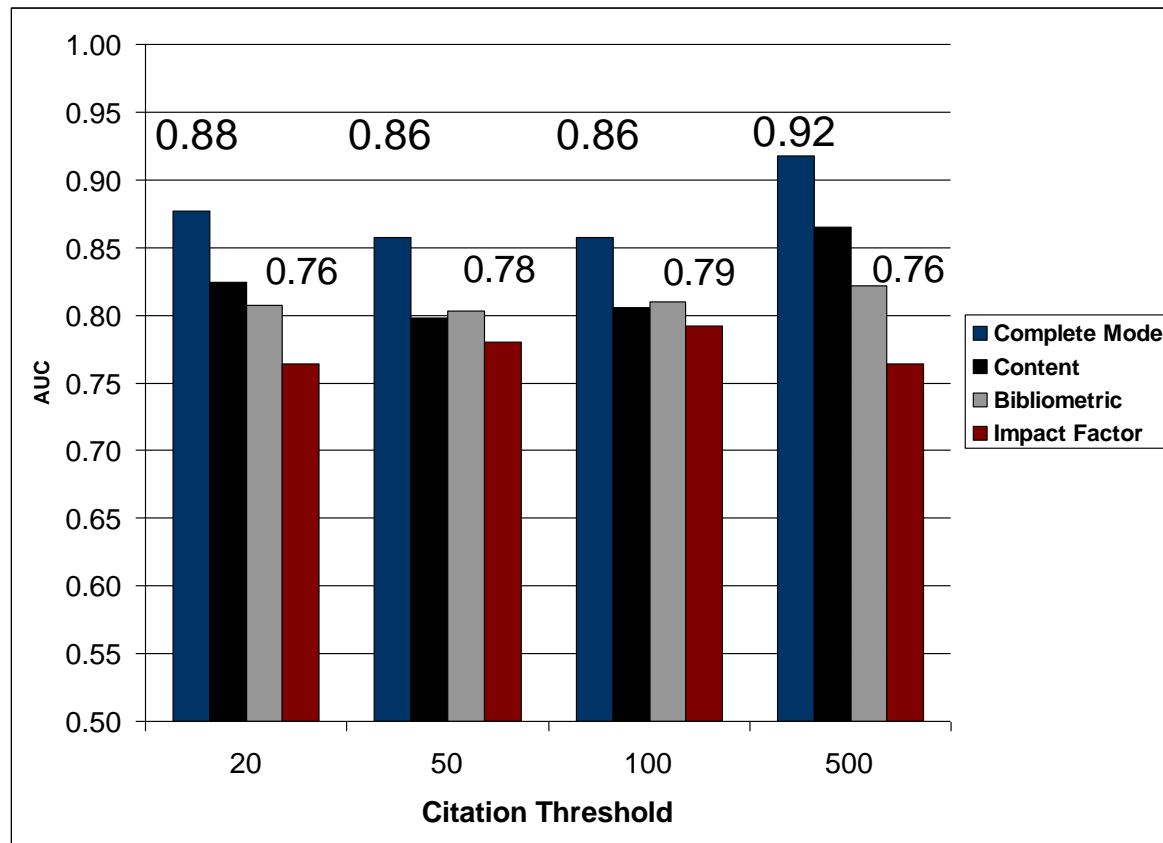
■ Prospective validation

- Trained models on 1991-1992 articles
- Evaluated performance on 1993-1994 articles



Results: Analysis of Feature Type

- Performance maximized with complete model
- Content, bibliometric features both important
- Impact Factor had worst performance



Results: Analysis of Influential Features

■ Feature Selection

- Markov Blanket: smallest set of features conditioned on which all remaining features are independent of the response variable
- Used HITON-PC algorithm
- Performance did not degrade with feature subset

■ Logistic Regression on selected features

- Estimate effect of each variable on the conditional probability of the response variable while controlling for other features

Results: Analysis of Influential Features

- Some topics indicated high citation rates
- Citation history of authors, Impact Factor were highly ranked for all thresholds
- Important content features changed for different thresholds

Example Features for Threshold 100

Feature	Logistic Regression Coefficient	P-value	Standard Error
First Author Citations [WOS]	5.75	0	1.47
Smoking:mortality [MeSH]	4.22	0.018	1.79
Journal Impact Factor [WOS]	3.32	0	0.18
Last Author Citations [WOS]	3.02	0.001	0.87
Birth Weight [MeSH]	2.95	0	0.77
Pilot Projects [MeSH]	-2.91	0.013	1.17
Autoantibodies:blood [MeSH]	2.78	0.001	0.81
Family Practice [MeSH]	-2.75	0.016	1.14
gy	2.65	0.006	0.96

A positive unit change in a regression coefficient β for a feature corresponds to e^β increase in the odds of exceeding the citation count threshold.

Focus 3: Classification of Citations

- Purpose: examine feasibility of automatically differentiating between instrumental and non-instrumental citations
- Benefits
 - Add functionality to citation indexers
 - Improve Impact Factor, Citation count by ignoring non-essential citations
- Previous approaches: manually generated rules
 - Teufel (2006): cue phrases, part-of-speech recognizer
 - Mercer (2004): cue phrases, grammar-like parsing rules

Definition of Instrumental Citation

- Hypothesis motivated by cited work
- Cited work necessary to complete citing work

A Tobacco-Specific Lung Carcinogen in the Urine of Men Exposed to Cigarette Smoke

Stephen S. Hecht, Steven G. Carmella, Sharon E. Murphy, Shobha Akerkar, Klaus D. Brunnemann, and Dietrich Hoffmann

NNK is a powerful pulmonary carcinogen, inducing predominantly adenocarcinomas in the lungs of rats, mice, and hamsters regardless of the route of administration^{5,6,7}. The lowest total doses required



Induction of Lung and Exocrine Pancreas Tumors in F344 Rats by Tobacco-specific and *Areca*-derived *N*-Nitrosamines^{1,2}

Abraham Rivenson, Dietrich Hoffmann,³ Bogdan Prokopczyk, Shantu Amin, and Stephen S. Hecht

exposure levels. These results support our hypothesis that NNK is a causative agent for cancers induced in humans by tobacco smoke. In this context one needs to consider the levels of

Predictive Features, Response Variable

Feature	Source	Representation			
Article title	MEDLINE	~20000 features after processing	Content Features		
Article abstract					
MeSH terms					
Citation text					
Number of times cited in Introduction	Article Full-text		Bibliometric Features		
Number of times cited in Methods					
Number of times cited in Results					
Number of times cited in Discussion					
Citation count of reference					
Number of articles for first author				Web of Science: not publically available, has to be extracted	1 integer value
Number of citations for first author					
Number of articles for last author					
Number of citations for last author					
Number of authors				www.arwu.org	4 values
Number of institutions					
Quality of first author's institution	Manually Labeled	1 binary value			
Instrumental Label					

Corpus Construction



The NEW ENGLAND
JOURNAL of MEDICINE

[FREE NEJM E-TOC](#) | [HOME](#) | [SUBSCRIBE](#) | [CURRENT ISSUE](#) | [PAST ISSUES](#) | [COLLECTIONS](#)

[Sign in](#) | [Get NEJM's E-Mail Table of Contents — Free](#) | [Subscribe](#)

ORIGINAL ARTICLE

[◀ Previous](#) | [Volume 329:1770-1776](#) | [December 9, 1993](#) | [Number 24](#) | [Next ▶](#)

Escalated as Compared with Standard Doses of Doxorubicin in BACOP Therapy for Patients with Non-Hodgkin's Lymphoma

*Ralph M. Meyer, Ian C. Quirt, Jamey R. Skillings, M.C. Cripps, Vivien Bramwell,
Brian H. Weinerman, Mary K. Gospodarowicz, Bruce F. Burns, Ann Marie Sargeant,
Lois E. Shepherd, Benny Zee, and William M. Hryniuk*

ABSTRACT

Background and Methods In 1981 the Clinical Trials Group of the National Cancer Institute of Canada completed a pilot study in patients with advanced-stage non-Hodgkin's lymphoma with aggressive tumor histology. That study demonstrated the potential efficacy of escalating the dose of doxorubicin used in

THIS ARTICLE

▶ [Abstract](#)

TOOLS & SERVICES

- ▶ [Add to Personal Archive](#)
- ▶ [Add to Citation Manager](#)
- ▶ [Notify a Friend](#)
- ▶ [E-mail When Cited](#)

1. **Download full text** of articles
 - New England Journal of Medicine
 - Internal Medicine articles from 1993, 1994

Corpus Construction

Most multidrug regimens used in treating non-Hodgkin's lymphoma with aggressive histologic features include an anthracycline drug, such as doxorubicin^{3,4,5,6,7,8}. The importance of including an anthracycline was demonstrated in a randomized trial comparing a regimen of cyclophosphamide, doxorubicin, vincristine, and prednisone (CHOP) plus bleomycin with a regimen of cyclophosphamide, vincristine, and prednisone plus bleomycin¹³. In two other randomized trials, regimens that included doxorubicin were superior to those that did not^{14,15}.

2. For 3 randomly selected references:

- Parse **citation text**
- Count **number of times cited** in each section
- Manually **label instrumental citations**

Corpus Construction

Title

1: J Infect Dis. 2002 Mar 1;185(5):650-6. Epub 2002 Feb 14.

[Related Articles](#), [Links](#)

The University of
Chicago Press

The clinical significance of cerebrospinal fluid levels of kynurenine pathway metabolites and lactate in severe malaria

Medana IM, Hien TT, Day NP, Phu NH, Mai NT, Chu'ong LV, Chau TT, Taylor A, Salahifar H, Stocker R, Smythe G, Turner GD, Farrar J, White NJ, Hunt NH.

Nuffield Department of Clinical Laboratory Sciences, Oxford-Wellcome Centre for Tropical and Infectious Diseases

Abstract

A retrospective study of 261 Vietnamese adults with severe malaria was conducted to determine the relationship between cerebrospinal fluid (CSF) levels of metabolites of the kynurenine pathway, the incidence of neurologic complications, and the disease outcome. Three metabolites were measured: the excitotoxin quinolinic acid (QA); the protective receptor antagonist kynurenic acid (KA); and the proinflammatory mediator picolinic acid (PA). These measurements were related prospectively to CSF lactate levels. QA and PA levels were elevated, compared with those of controls. There was no difference in the levels of KA between these groups. Although >40% of malaria patients had QA CSF concentrations in the micromolar range, there was no association with convulsions or depth of coma. Levels of QA and PA were associated significantly with death, but a multivariate analysis suggested that these elevations were a consequence of impaired renal function. CSF lactate remained an independent and significant predictor of poor outcome.

Publication Types:

- Clinical Trial
- Randomized Controlled Trial

MeSH terms

MeSH Terms:

- Malaria, Cerebral/cerebrospinal fluid*
- Malaria, Cerebral/drug therapy
- Malaria, Cerebral/parasitology

PMID: 11865422 [PubMed - indexed for MEDLINE]

3. Download content features from MEDLINE using Python scripts

Corpus Construction

ISI Web of KnowledgeSM Take the next step

All Databases Select a Database Web of Science Additional Resources

Search Search History Marked List (0)

ALL DATABASES

<< Back to results list Record 1 of 2 >>

The clinical significance of cerebrospinal fluid levels of kynurenine pathway metabolites and lactate in severe malaria

find it NCBI Print Email Add to Marked List Save to EndNote@Web

Holdings Go Save to EndNote, RefMan, ProCite more options

Author(s): Medana IM, Nien TT, Day NP, Phu NH, Mai NTH, Chu'ong LV, Chau TTH, Taylor A, Salahifar H, Stocker R, Smythe G, Turner GDH, Farrar J, White NJ, Hunt NH

Source: JOURNAL OF INFECTIOUS DISEASES Volume: 185 Issue: 5 Pages: 650-656 Published: MAR 1 2002

Times Cited: 25 References: 28 Citation Map

Results Author=(Medana IM) Timespan=All Years

Results: 22 Page 1 of 3 Go Sort by: Publication Date

Refine Results Search within results for

General Categories (Refine) SCIENCE & TECHNOLOGY (22) SOCIAL SCIENCES (1) more options / values

Subject Areas (Refine) NEUROSCIENCES & NEUROLOGY (20) PARASITOLOGY (19) IMMUNOLOGY (15) BIO-CHEMISTRY & MOLECULAR BIOLOGY (13) INFECTIOUS DISEASES (12) more options / values

Document Types Authors Source Titles Publication Years Languages

1. Title: Host vascular endothelial growth factor is trophic for Plasmodium falciparum-infected red blood cells
Author(s): Sachanonta, N, Medana, IM, Roberts, R, et al
Source: ASIAN PACIFIC JOURNAL OF ALLERGY AND IMMUNOLOGY Volume: 26 Issue: 1 Pages: 37-45 Published: 2008 Times Cited: 0 Find it

2. Title: Fatal cerebral malaria: distinct microvascular pathologies in children and adult patients
Author(s): Wassmer, SC, Medana, IM, Turner, GDH, et al
Source: INTERNATIONAL JOURNAL FOR PARASITOLOGY Volume: 38 Pages: S44-S44 Published: 2008

Citing Articles

Title: The clinical significance of cerebrospinal fluid levels of kynurenine pathway metabolites and lactate in severe malaria
Author(s): Medana, IM
Source: JOURNAL OF INFECTIOUS DISEASES Volume: 185 Issue: 5 Pages: 650-656 Published: MAR 1 2002
Citation Map

The above article has been cited by the articles listed below
Note: The Times Cited count is calculated across all Web of Science editions. More information.

Results: 25 Page 1 of 3 Go Sort by: Latest Date

Refine Results Search within results for

Subject Areas (Refine) PARASITOLOGY (2) IMMUNOLOGY (4) NEUROSCIENCES (4) INFECTIOUS DISEASES (3) BIO-CHEMISTRY & MOLECULAR BIOLOGY (2) more options / values

Document Types (Refine) ARTICLE (11) REVIEW (3) PROCEEDINGS PAPER (4) MEETING ABSTRACT (2) more options / values

Authors Source Titles Publication Years Institutions Languages Countries/Territories

1. Title: Tryptophan in physiology and pathophysiology
Author(s): Hunt NH-Source: FREE RADICAL RESEARCH Volume: 42 Pages: S25-S25 Supplement: Suppl. 1 Published: JUL 2008 Times Cited: 0 Find it

2. Title: In-hospital risk estimation in children with Malaria - Early predictors of morbidity and mortality
Author(s): Winler AS, Sainbohor G, Heibok R, et al-Source: JOURNAL OF TROPICAL PEDIATRICS Volume: 54 Issue: 3 Pages: 184-191 Published: JUN 2008 Times Cited: 0 Find it

3. Title: Immunosuppression routed via the kynurenine pathway: A biochemical and pathophysiologic approach
Author(s): Gonzalez A, Vero N, Alegre E, et al-Source: ADVANCES IN CLINICAL CHEMISTRY, VOL 45 Book Series: ADVANCES IN CLINICAL CHEMISTRY Volume: 45 Pages: 155-197 Published: 2008 Times Cited: 1 Find it

4. Title: Characterization of an indoleamine 2,3-dioxygenase-like protein found in humans and mice
Author(s): Ball HJ, Sanchez-Perez A, Weller S, et al-Source: GENE Volume: 396 Issue: 1 Pages: 203-213 Published: JUL 1 2007 Times Cited: 12 Find it

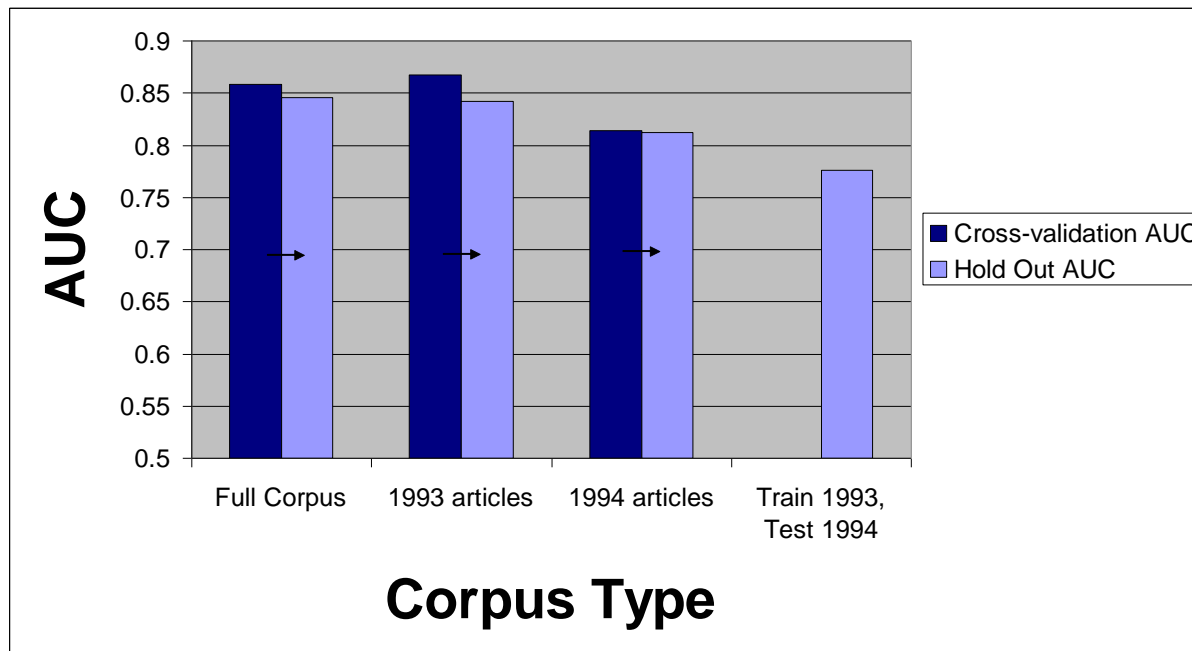
- 4. Download bibliometric features from Web of Science
- Simulate user session with Python scripts
- Manually extract features for difficult cases

Experimental Design

- Document Representation
 - Bag of words
 - Porter stemming
 - Term weighting: log frequency with redundancy
- Learning Method
 - Support Vector Machine (SVM) models
- Model Selection
 - 5-fold nested cross-validation
 - Performance metric: AUC

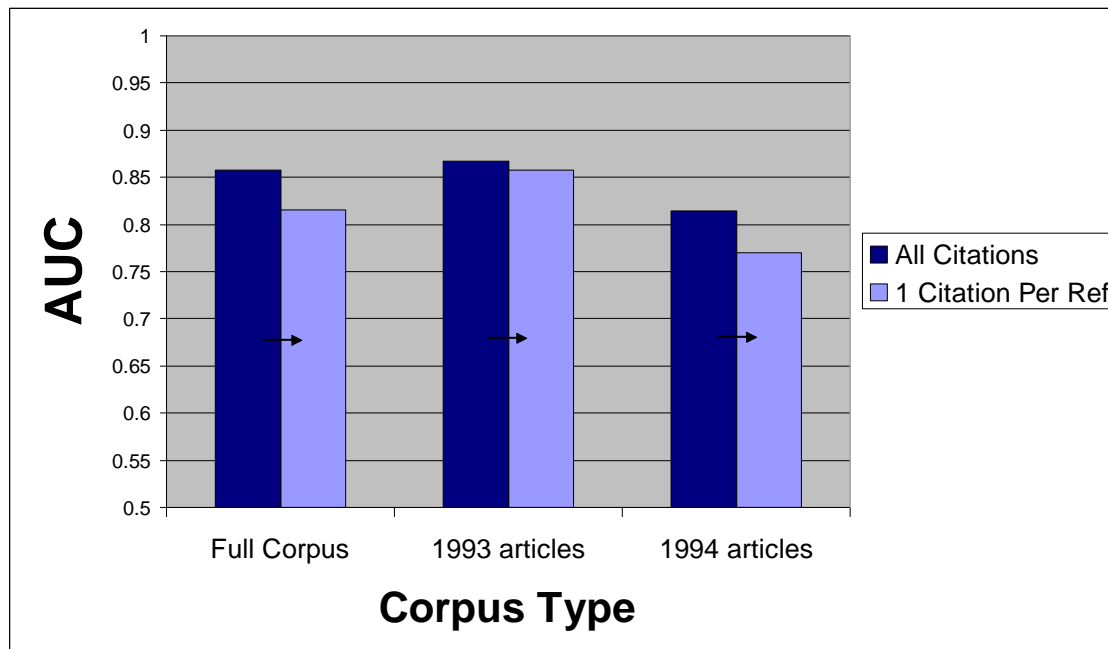
Results

- Cross-validation: accurate classification possible
- Repeated experiments with hold out test sets
 - Exclude test set before model learning
 - More robust estimate of generalizability
- Models may be time-dependent



Results

- Citations to same reference not independent
- Performed analysis after restricting corpus to one citation per reference
- Results still showed relatively accurate classification



Analysis of Influential Features

- Influential features identified with feature selection
 - Markov Blanket induction
 - Logistic Regression

Feature	Example Features		
	Logistic Regression Coefficient	P-value	Standard Error
Number of times cited in introduction [WOS]	5.65	0.00	0.70
von	-3.42	0.01	1.31
mammographi	-2.90	0.02	1.25
Cytarabine[MeSH]	-2.70	0.00	0.80
Arrhythmias, Cardiac [MeSH]	-2.43	0.01	0.97
complex	-2.38	0.00	0.71
eject	-2.20	0.00	0.73
visual	-1.97	0.00	0.65
underestim	-1.89	0.01	0.69

Discussion

- Understand if factors are causative or unrelated to quality
- Investigate how robust prediction models are to manipulation
- Improve performance of models
- Create modified versions of Impact Factor, citation count, prediction models using only instrumental citations
- Evaluate performance and impact in real world

Conclusion

- Topic-sensitivity
 - Demonstrates topic-sensitivity of Impact Factor, Clinical Query Filters, and PageRank
 - Proposes alternative methods that are stable over topics
- Citation Count Prediction
 - Demonstrates that citation count prediction is possible using information at publication time
 - Provides insight into the factors affecting citation behavior
- Automatic Classification of Instrumental Citations
 - Feasible to automatically identify instrumental citations
 - Results may potentially improve existing citation metrics
- Opens door to many real world extensions and future work

Acknowledgements

■ Committee Members

- Constantin Aliferis
- Cindy Gadd
- Nunzia Giuse
- Daniel Masys
- Lily Wang

■ Alex Statnikov, Yindalon Aphinyanaphongs