

# Pattern recognition models to predict citation count: Main Idea

1. Build **training corpora** with bibliometric and content features that are predictive of citation count

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

**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

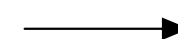
MeSH Terms:

- Malaria, 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, and MeSH terms)

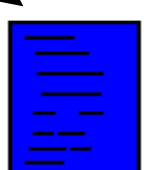
Labeled Examples



```

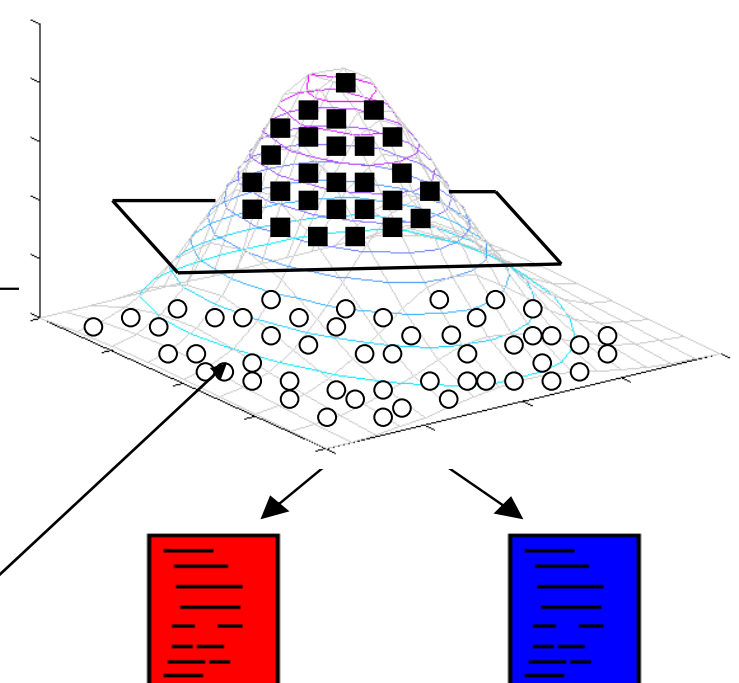
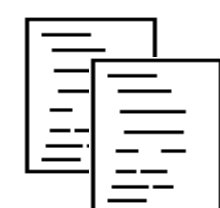
MESH (LFL:LowType)
1.  CH_CITITERBANDP
2.  CH_CITITERBANDP
3.  CH_CITITERBANDP
4.  CH_CITITERBANDP
5.  CH_CITITERBANDP
6.  CH_CITITERBANDP
7.  CH_CITITERBANDP
8.  CH_CITITERBANDP
9.  CH_CITITERBANDP
10. CH_CITITERBANDP
11. CH_CITITERBANDP
12. CH_CITITERBANDP
13. CH_CITITERBANDP
14. CH_CITITERBANDP
15. CH_CITITERBANDP
16. CH_CITITERBANDP
17. CH_CITITERBANDP
18. CH_CITITERBANDP
19. CH_CITITERBANDP
20. CH_CITITERBANDP
21. CH_CITITERBANDP
22. CH_CITITERBANDP
23. CH_CITITERBANDP
24. CH_CITITERBANDP
25. CH_CITITERBANDP
26. CH_CITITERBANDP
27. CH_CITITERBANDP
28. CH_CITITERBANDP
29. CH_CITITERBANDP
30. CH_CITITERBANDP
31. CH_CITITERBANDP
32. CH_CITITERBANDP
33. CH_CITITERBANDP
34. CH_CITITERBANDP
35. CH_CITITERBANDP
36. CH_CITITERBANDP
37. CH_CITITERBANDP
38. CH_CITITERBANDP
39. CH_CITITERBANDP
40. CH_CITITERBANDP
41. CH_CITITERBANDP
42. CH_CITITERBANDP
43. CH_CITITERBANDP
44. CH_CITITERBANDP
45. CH_CITITERBANDP
46. CH_CITITERBANDP
47. CH_CITITERBANDP
48. CH_CITITERBANDP
49. CH_CITITERBANDP
50. CH_CITITERBANDP
51. CH_CITITERBANDP
52. CH_CITITERBANDP
53. CH_CITITERBANDP
54. CH_CITITERBANDP
55. CH_CITITERBANDP
56. CH_CITITERBANDP
57. CH_CITITERBANDP
58. CH_CITITERBANDP
59. CH_CITITERBANDP
60. CH_CITITERBANDP
61. CH_CITITERBANDP
62. CH_CITITERBANDP
63. CH_CITITERBANDP
64. CH_CITITERBANDP
65. CH_CITITERBANDP
66. CH_CITITERBANDP
67. CH_CITITERBANDP
68. CH_CITITERBANDP
69. CH_CITITERBANDP
70. CH_CITITERBANDP
71. CH_CITITERBANDP
72. CH_CITITERBANDP
73. CH_CITITERBANDP
74. CH_CITITERBANDP
75. CH_CITITERBANDP
76. CH_CITITERBANDP
77. CH_CITITERBANDP
78. CH_CITITERBANDP
79. CH_CITITERBANDP
80. CH_CITITERBANDP
81. CH_CITITERBANDP
82. CH_CITITERBANDP
83. CH_CITITERBANDP
84. CH_CITITERBANDP
85. CH_CITITERBANDP
86. CH_CITITERBANDP
87. CH_CITITERBANDP
88. CH_CITITERBANDP
89. CH_CITITERBANDP
90. CH_CITITERBANDP
91. CH_CITITERBANDP
92. CH_CITITERBANDP
93. CH_CITITERBANDP
94. CH_CITITERBANDP
95. CH_CITITERBANDP
96. CH_CITITERBANDP
97. CH_CITITERBANDP
98. CH_CITITERBANDP
99. CH_CITITERBANDP
100. CH_CITITERBANDP
    
```

Labeled

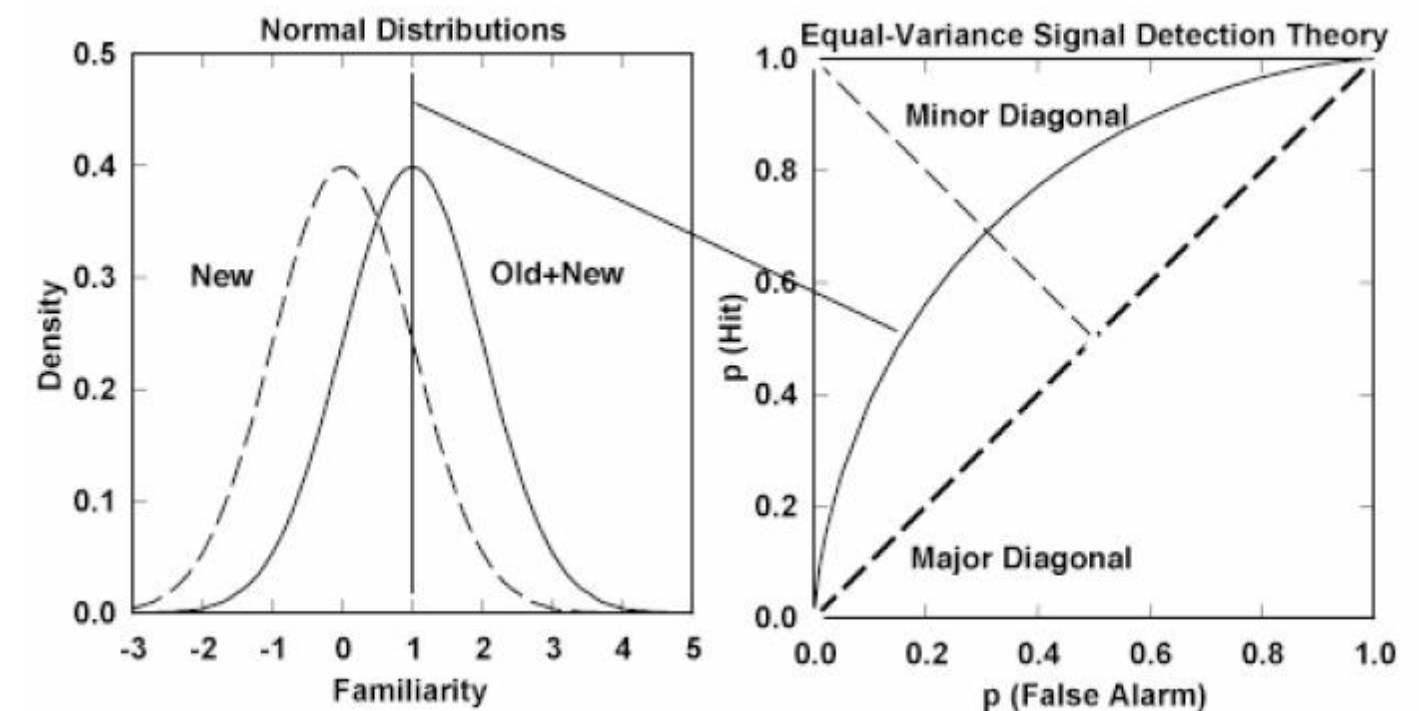
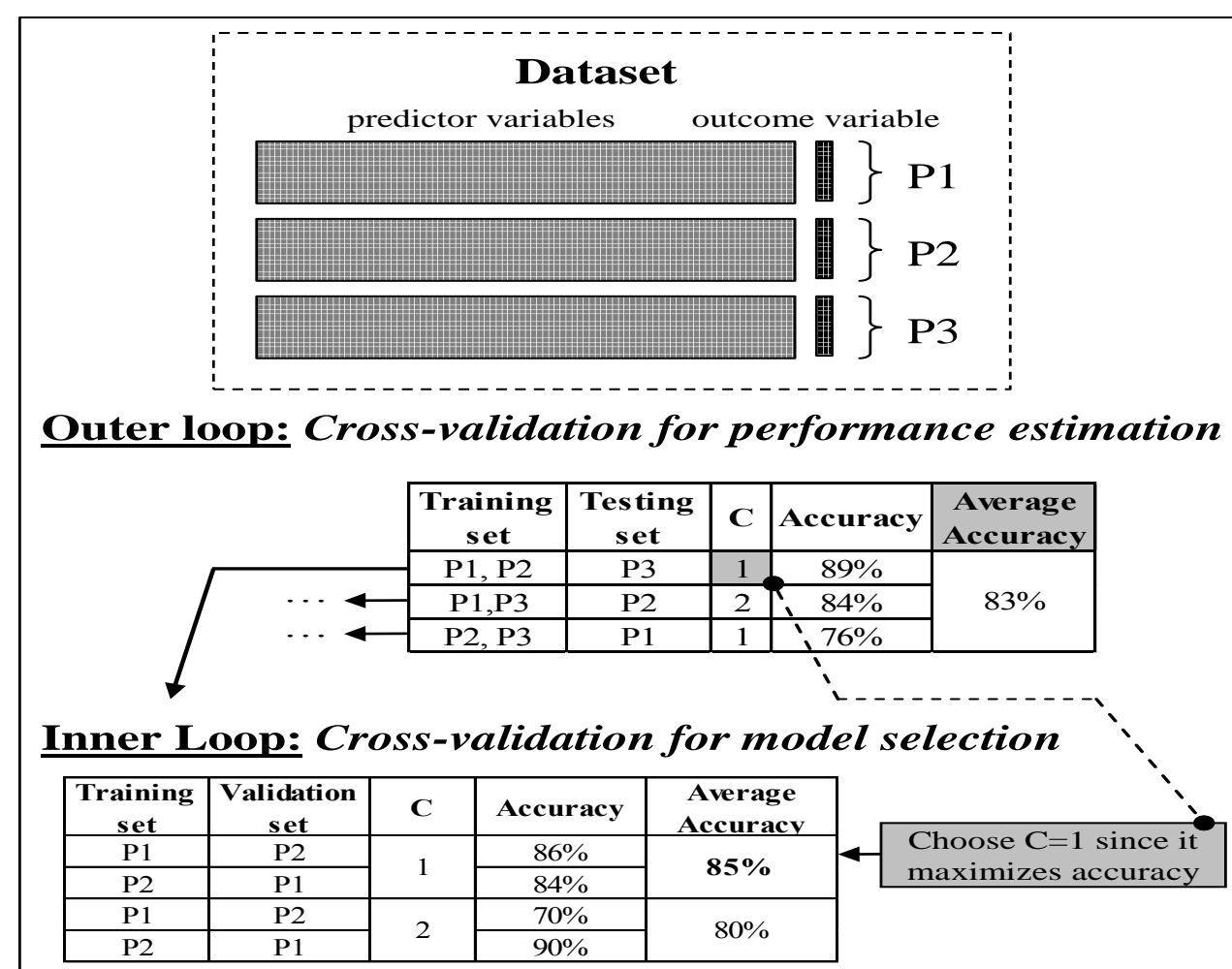


3. Train **models** that capture implicit quality criteria and can handle high-dimensional text data

Unseen Examples

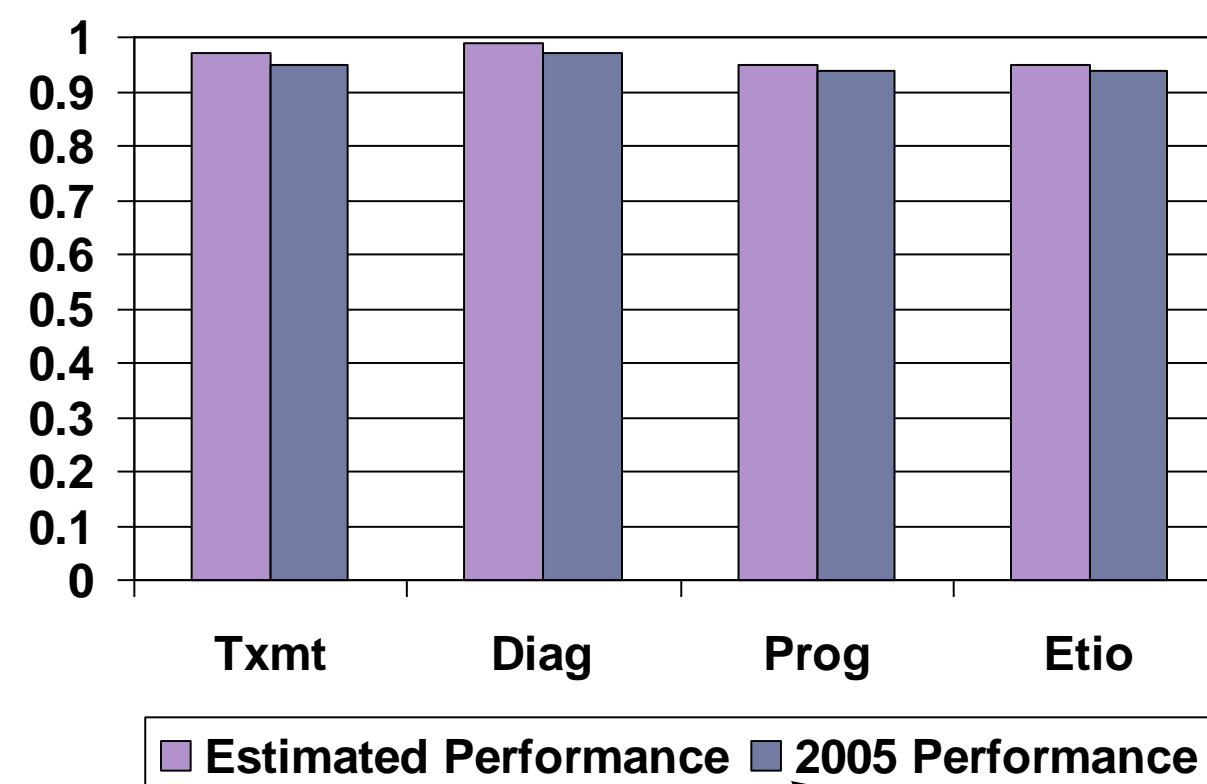


# Pattern recognition models to predict citation count: Main Idea Continued



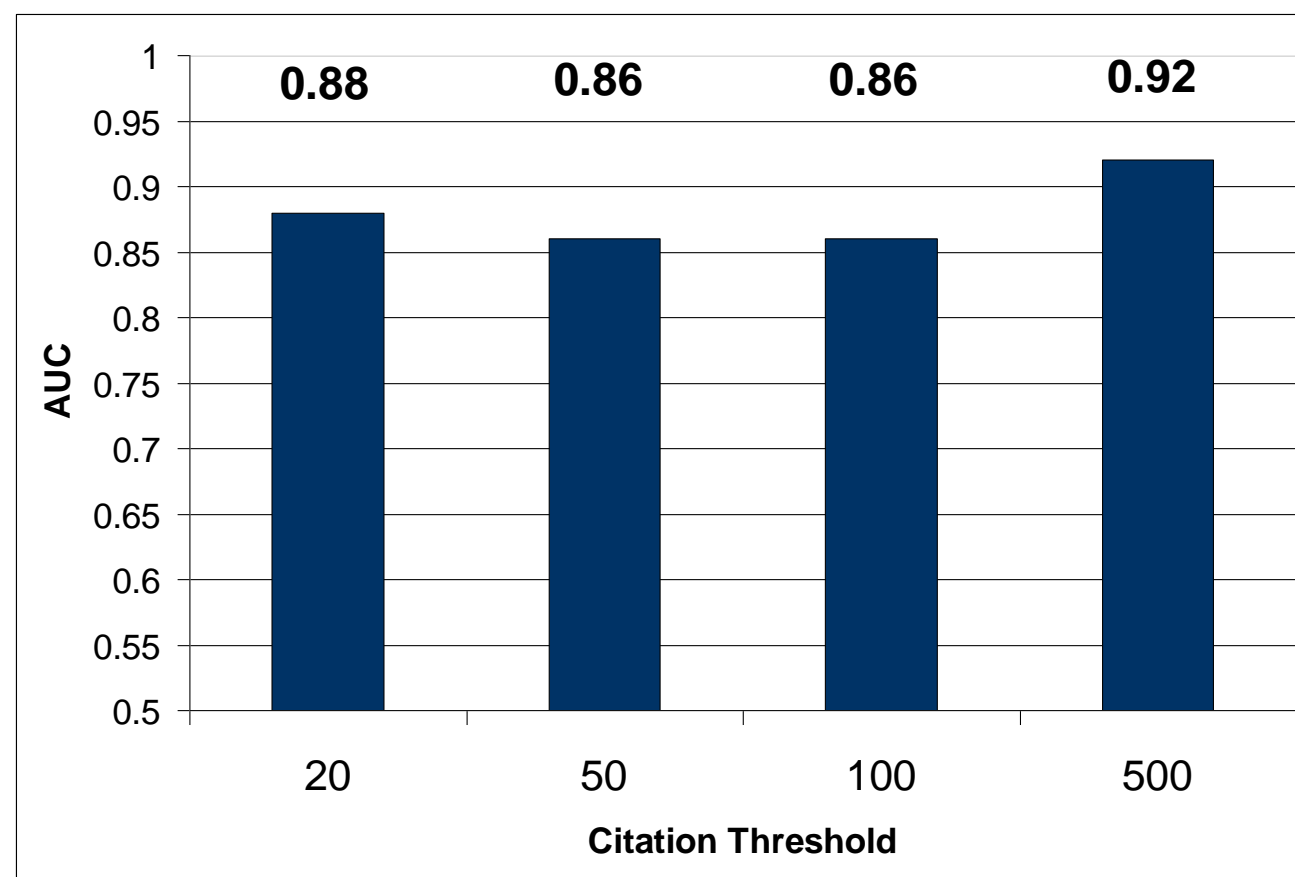
## 4. Evaluate models' performances

- with **nested cross-validation** to estimate error
- use **AUC** as performance metric

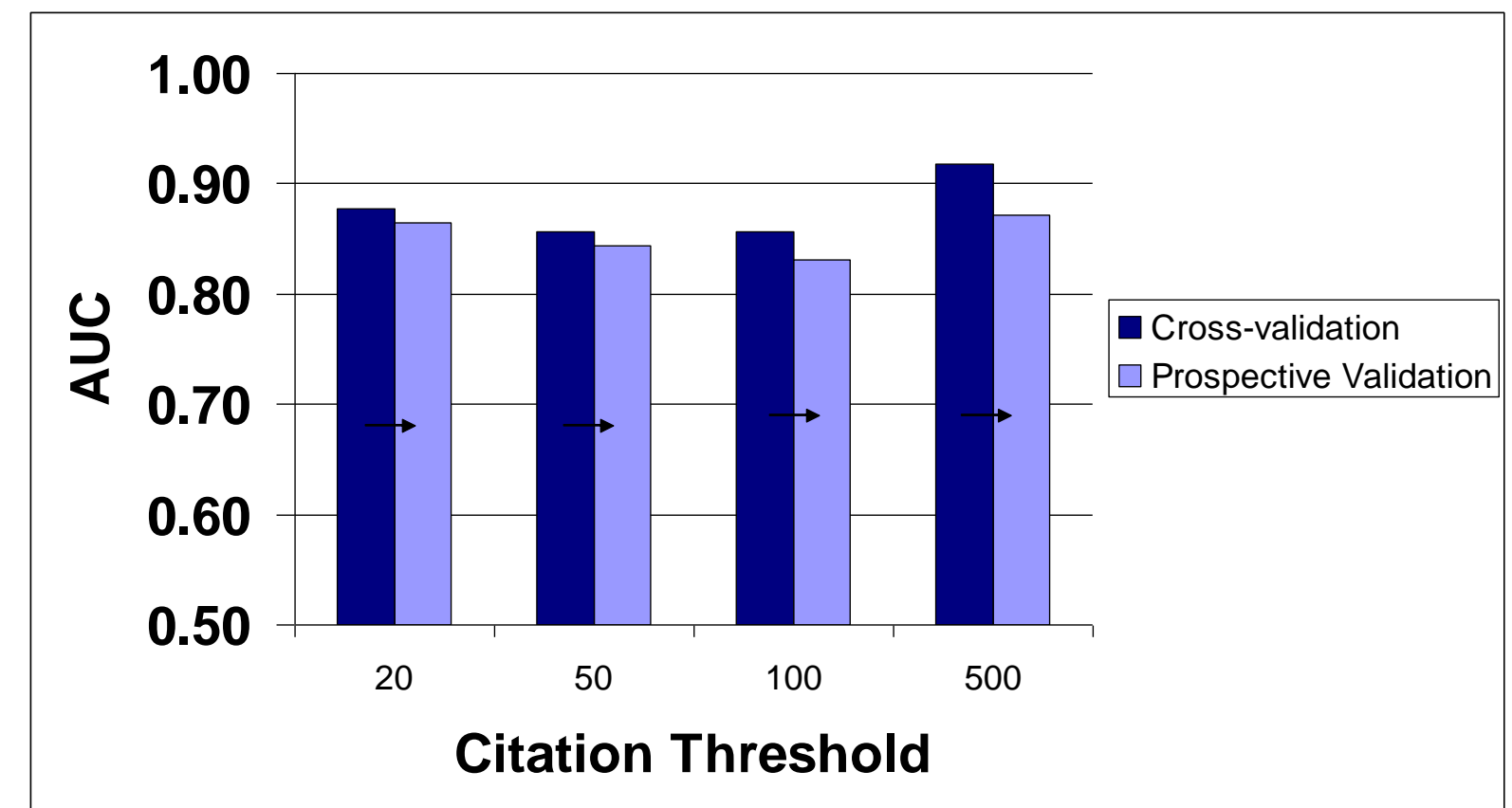


## 5. Evaluate performance **prospectively** & compare to prior cross-validation estimates

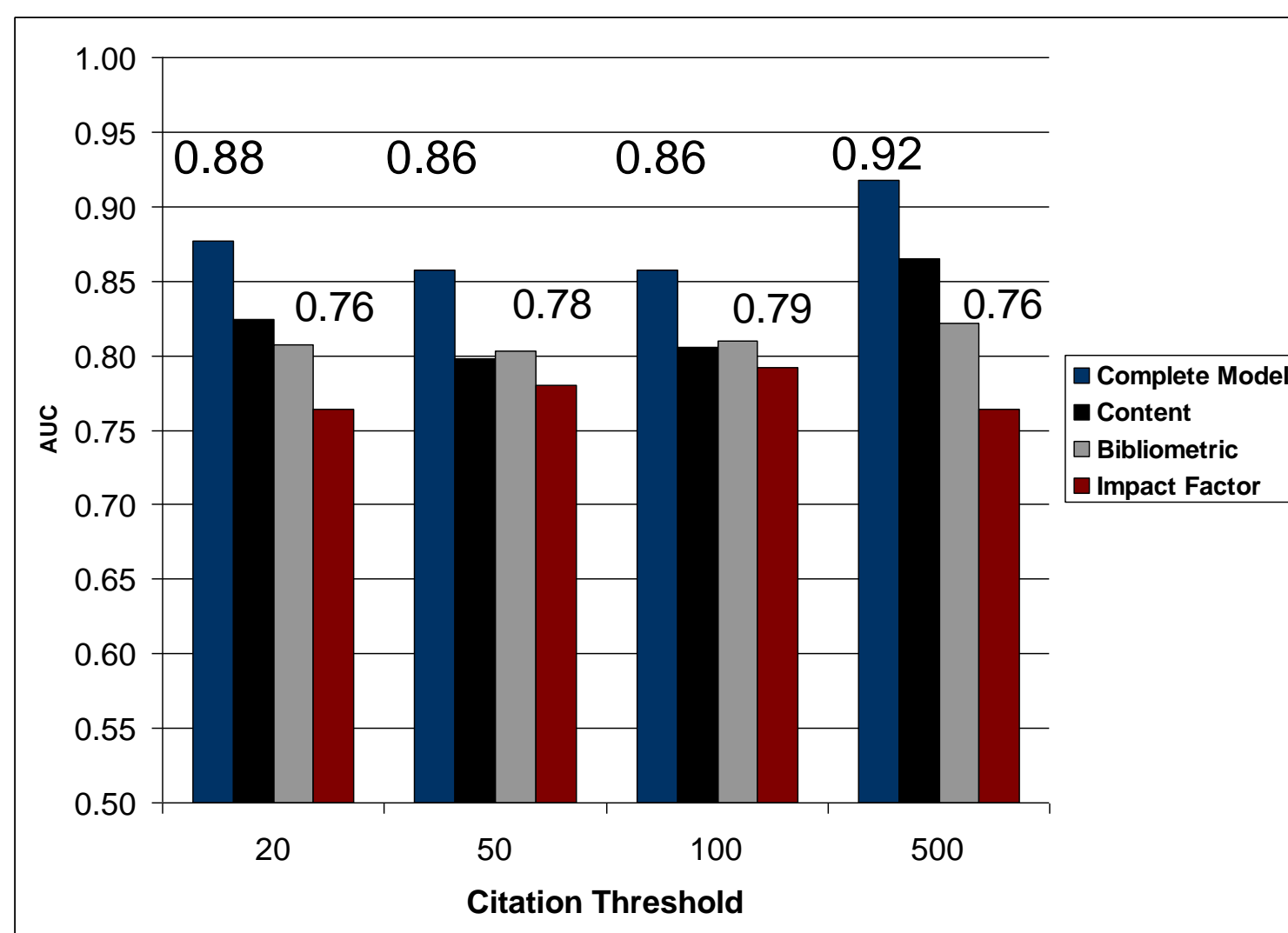
# Pattern recognition models to predict citation count: Some Notable Results



1. Possible to build models with only information at publication time that result in **high predictivity**



2. Independent prospective validation results show that models are not overfitted



3. Performance of models trained on feature subsets shows that both types of features are needed for best performance

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

4. Important features identified by performing feature selection and logistic regression

# Future Work

---

**The results so far are very exciting and clearly open many future directions for research:**

- Develop better understanding of whether predictive factors are **causative or confounders** of hidden variables
- Determine **if models can be manipulated** (i.e. authors can take advantage of knowledge gained from models to increase citability without improving quality of work)
- **Improve model predictivity** by including additional features, using different data sources
- Deploy the models in **real world applications** and rebuild them for more topics and journals

**We have also developed machine learning models that can accurately characterize whether a citation is essential.**

- Citation counts can be **adjusted by discarding non-essential citations** that are not important to the citing papers (many highly-cited papers receive numerous citations due to factors other than quality)
- Rebuild citation prediction models to **predict essential citation counts**