What Predicts Media Coverage of Health Science Articles?

Byron C Wallace△, Michael J Paul‡ & Noémie Elhadad †

△ University of Texas at Austin; byron.wallace@utexas.edu
‡ Johns Hopkins University
† Columbia University

"This magazine says we can lose 50 pounds in a week by eating chocolate cake three times a day. Finally, a diet that makes sense!"

Health science & the media

- Timely and accurate reporting is critical for communicating important new findings to the public
- Journalists are implicitly tasked with selecting a handful of ‘newsworthy’ articles from the huge volume published every day
- We don’t know much about how these decisions are made

Aim

- Can we predict which published health science articles will receive media attention?
  - If so, what are the most discriminative features?

- *a priori* we might expect such articles to contain words involving weight-loss, etc.
  - other expectations?

Corpus construction

- It's easy to identify *positive* examples

Obese mothers have babies with more belly fat, study finds

BY SHEREEN LEHMAN

NEW YORK | Tue Jun 24, 2014 3:10pm EDT

RELATED TOPICS

Health »

(Reuters Health) - Babies of obese mothers tend to be born with more fat, especially around their middles, than babies with leaner mothers, according to a new study.

...

SOURCE: bit.ly/1iBKiaR Acta Paediatrica, online June 18, 2014.

Obese mothers have babies with more belly fat, study finds

BY SHEREEN LEHMAN

NEW YORK | Tue Jun 24, 2014 3:10pm EDT

RELATED TOPICS

Health »

(Reuters Health) - Babies of obese mothers tend to be born with more fat, especially around their middles, than babies with leaner mothers, according to a new study.

SOURCE: bit.ly/1iBKiaR Acta Paediatrica, online June 18, 2014

Corpus construction

- But what constitute *negative* examples?

Matched sampling

- Strategy: identify articles with characteristics such that they should have had equal chance of being ‘picked up’ by the media but (very probably) were not

- Specifically: sample a set of ‘negative’ articles for each ‘positive’ article that were published in the same journal and in the same year

Matched sampling

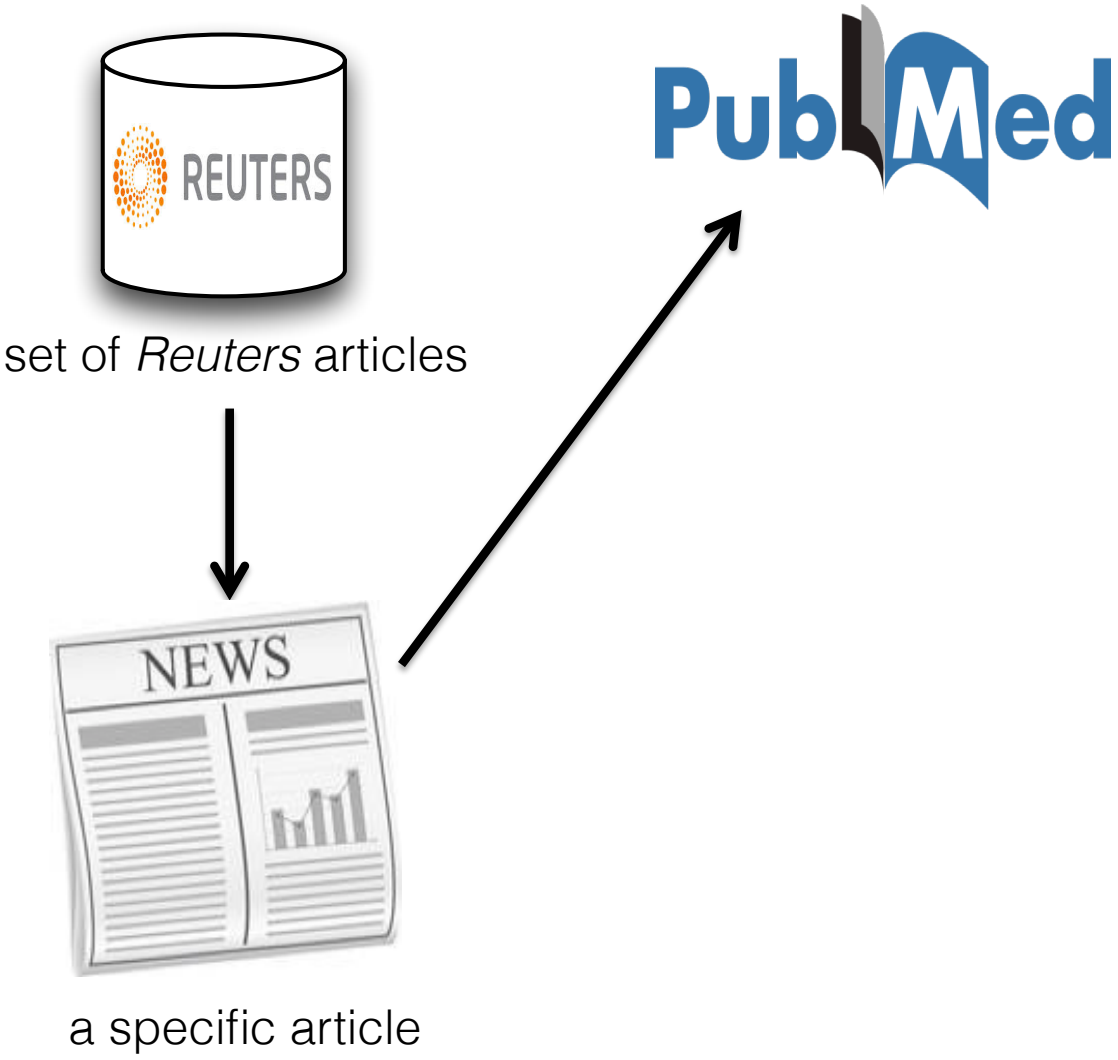
set of *Reuters* articles

Matched sampling

set of *Reuters* articles

a specific article

Matched sampling



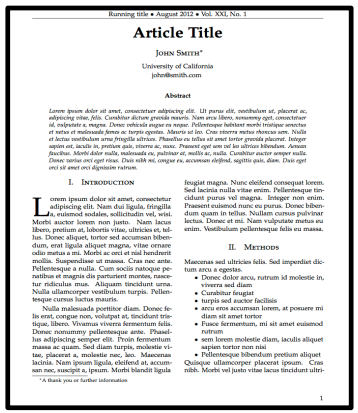
set of *Reuters* articles

a specific article

Matched sampling

set of *Reuters* articles

a specific article



Running title • August 2012 • Vol. XXI, No. 1

Article Title

JOHN SMITH*

University of California
john@smith.com

Abstract

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

I. INTRODUCTION

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

II. METHODS

Maecenas sed ultricies felis. Sed imperdiet dictum arcu a egestas.

- Donec dolor arcu, rutrum id molestie in, viverra sed diam
- Curabitur feugiat
- turpis sed auctor facilisis
- arcu eros accumsan lorem, at posuere mi diam sit amet tortor
- Fusce fermentum, mi sit amet euismod rutrum
- sem lorem molestie diam, iaculis aliquet sapien tortor non nisi
- Pellentesque bibendum pretium aliquet

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultri-

*A thank you or further information

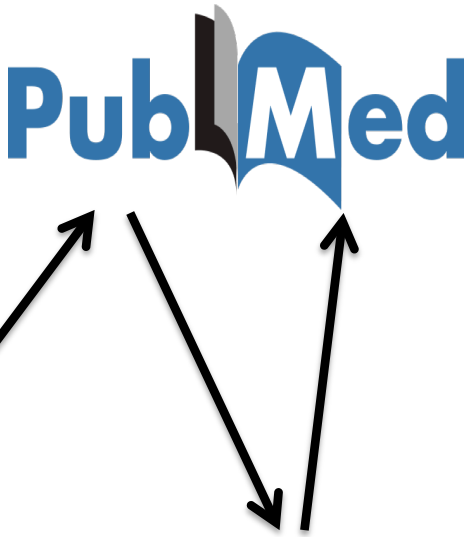
1

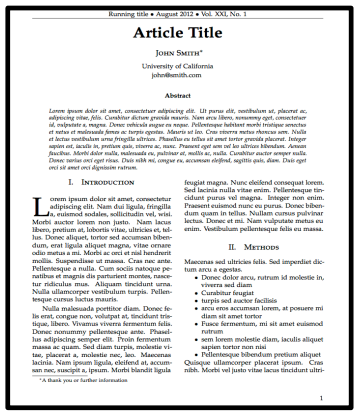
published scientific article

Matched sampling

set of *Reuters* articles

a specific article




Running title • August 2012 • Vol. XXI, No. 1

Article Title

John Smith*

University of California

john@smith.com

Abstract

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

I. Introduction

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

II. Methods

Maecenas sed ultricies felis. Sed imperdiet dictum arcu a egestas.

- Donec dolor arcu, rutrum id molestie in, viverra sed diam
- Curabitur feugiat
- turpis sed auctor facilisis
- arcu eros accumsan lorem, at posuere mi diam sit amet tortor
- Fusce fermentum, mi sit amet euismod rutrum
- sem lorem molestie diam, iaculis aliquet sapien tortor non nisi
- Pellentesque bibendum pretium aliquet

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultri-

*A thank you or further information

1

published scientific article

Matched sampling

set of *Reuters* articles

a specific article

published scientific article

matched scientific articles

Corpus

- We downloaded 1,343 (‘positive’) Reuters piece / article pairs
  - Reuters articles were published between 1/2012 and 9/2014

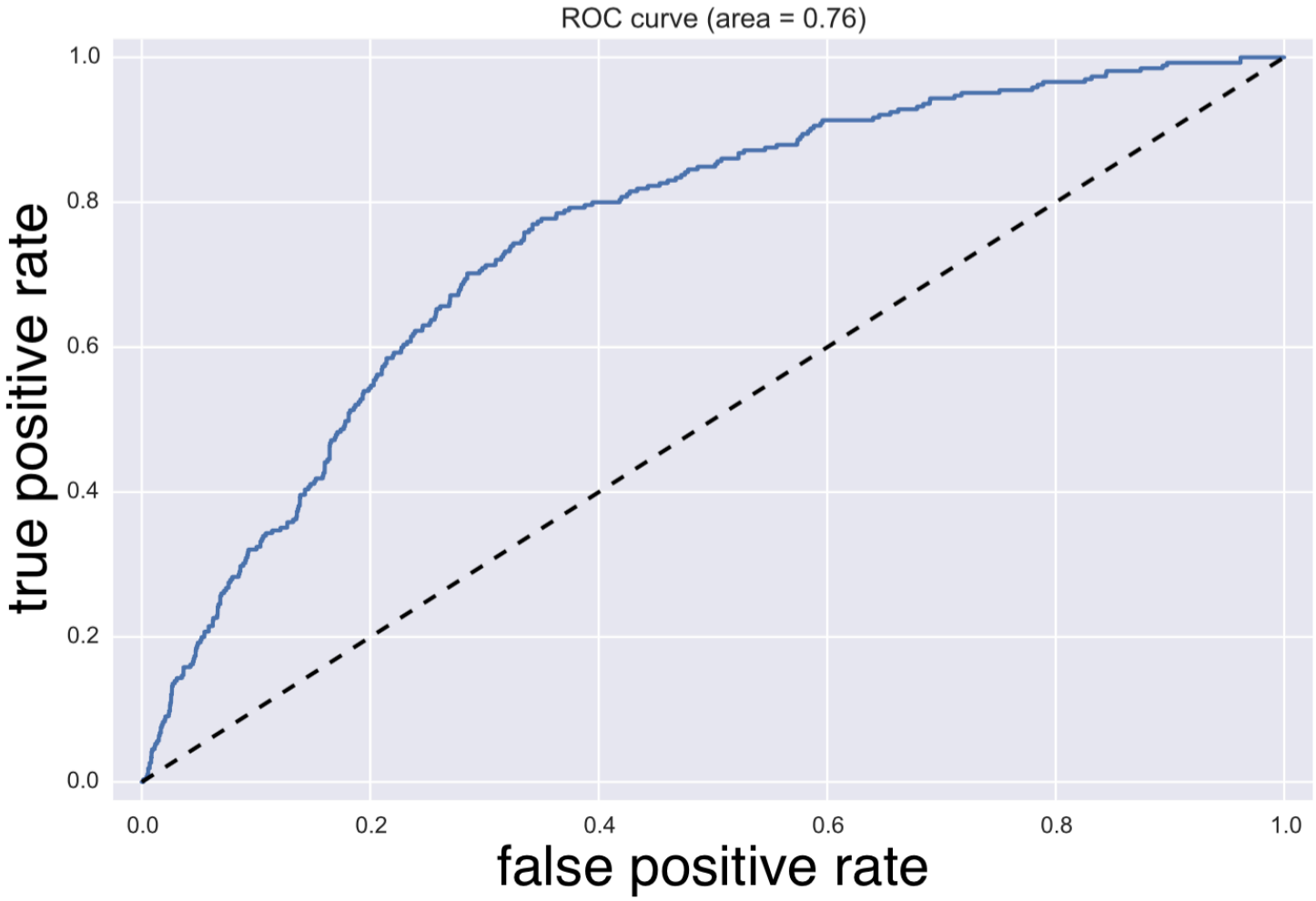
- For each article, we sampled up to 20 ‘matched’ articles published in the same journal and in the same year, but not covered by Reuters
  - Heuristically filtered to attempt to include only full-length original research articles
  - In total: 27,567 articles

- Our data is available @ https://github.com/bwallace/w3phi-2015

Learning

- Standard text classification pipeline
- Uni- and bi-gram Bag-of-Words representation encoding title, abstract and MeSH terms
  - English stopwording
  - Kept tokens observed in >= 100 articles
  - 14,614 features in total
- Logistic regression with a squared L2 norm penalty for regularization
  - We tuned the regularization parameter on the train set to maximize

Can we predict media coverage?



10-fold cross validation: average AUC 0.76, range (0.75, 0.811)

So, yes.
But what words correlate with media coverage?

Features that predict media coverage

negative		*positive*	
-1.102	patients	0.617	exercise
-0.507	clinical	0.610	mh-data
-0.457	2012	0.604	mh-numerical mh-data
-0.393	survival	0.586	mh-numerical
-0.364	therapy	0.536	intake
-0.337	complications	0.531	mh-adult
-0.323	surgical	0.508	cancer
-0.321	response	0.500	mh-effects
-0.308	plasma	0.492	years
-0.300	pediatric	0.461	mh-child
-0.285	diagnostic	0.459	virus
-0.281	imaging	0.455	mh-aged
-0.275	2013	0.443	smoking
-0.267	management	0.442	influenza
-0.265	expression	0.428	mh-female mh-humans
-0.258	factors	0.418	consumption
-0.252	outcomes	0.407	incident
-0.250	score	0.407	women
-0.248	range	0.407	weight
-0.246	treatment	0.391	mh-humans
-0.246	function	0.389	exposure
-0.245	diabetes	0.383	asd
-0.242	review	0.381	pregnancy
-0.236	os	0.380	year
-0.235	protein	0.378	mh-studies
-0.231	mice	0.372	mh-female
-0.231	serum	0.359	effect
-0.227	values	0.355	95
-0.223	model	0.354	age
-0.223	mm	0.349	mh-male

Features that predict media coverage

negative		*positive*	
-1.102	patients	0.617	exercise
-0.507	clinical	0.610	mh-data
-0.457	2012	0.604	mh-numerical mh-data
-0.393	survival	0.586	mh-numerical
-0.364	therapy	0.536	intake
-0.337	complications	0.531	mh-adult
-0.323	surgical	0.508	cancer
-0.321	response	0.500	mh-effects
-0.308	plasma	0.492	years
-0.300	pediatric	0.461	mh-child
-0.285	diagnostic	0.459	virus
-0.281	imaging	0.455	mh-aged
-0.275	2013	0.443	smoking
-0.267	management	0.442	influenza
-0.265	expression	0.428	mh-female mh-humans
-0.258	factors	0.418	consumption
-0.252	outcomes	0.407	incident
-0.250	score	0.407	women
-0.248	range	0.407	weight
-0.246	treatment	0.391	mh-humans
-0.246	function	0.389	exposure
-0.245	diabetes	0.383	asd
-0.242	review	0.381	pregnancy
-0.236	os	0.380	year
-0.235	protein	0.378	mh-studies
-0.231	mice	0.372	mh-female
-0.231	serum	0.359	effect
-0.227	values	0.355	95
-0.223	model	0.354	age
-0.223	mm	0.349	mh-male

“intuitively agreeable”

Features that predict media coverage

negative		*positive*	
-1.102	patients	0.617	exercise
-0.507	clinical	0.610	mh-data
-0.457	2012	0.604	mh-numerical mh-data
-0.393	survival	0.586	mh-numerical
-0.364	therapy	0.536	intake
-0.337	complications	0.531	mh-adult
-0.323	surgical	0.508	cancer
-0.321	response	0.500	mh-effects
-0.308	plasma	0.492	years
-0.300	pediatric	0.461	mh-child
-0.285	diagnostic	0.459	virus
-0.281	imaging	0.455	mh-aged
-0.275	2013	0.443	smoking
-0.267	management	0.442	influenza
-0.265	expression	0.428	mh-female mh-humans
-0.258	factors	0.418	consumption
-0.252	outcomes	0.407	incident
-0.250	score	0.407	women
-0.248	range	0.407	weight
-0.246	treatment	0.391	mh-humans
-0.246	function	0.389	exposure
-0.245	diabetes	0.383	asd
-0.242	review	0.381	pregnancy
-0.236	os	0.380	year
-0.235	protein	0.378	mh-studies
-0.231	mice	0.372	mh-female
-0.231	serum	0.359	effect
-0.227	values	0.355	95
-0.223	model	0.354	age
-0.223	mm	0.349	mh-male

What's up with these??

Features that predict media coverage

negative		*positive*	
-1.102	patients	0.617	exercise
-0.507	clinical	0.610	mh-data
-0.457	2012	0.604	mh-numerical mh-data
-0.393	survival	0.586	mh-numerical
-0.364	therapy	0.536	intake
-0.337	complications	0.531	mh-adult
-0.323	surgical	0.508	cancer
-0.321	response	0.500	mh-effects
-0.308	plasma	0.492	years
-0.300	pediatric	0.461	mh-child
-0.285	diagnostic	0.459	virus
-0.281	imaging	0.455	mh-aged
-0.275	2013	0.443	smoking
-0.267	management	0.442	influenza
-0.265	expression	0.428	mh-female mh-humans
-0.258	factors	0.418	consumption
-0.252	outcomes	0.407	incident
-0.250	score	0.407	women
-0.248	range	0.407	weight
-0.246	treatment	0.391	mh-humans
-0.246	function	0.389	exposure
-0.245	diabetes	0.383	asd
-0.242	review	0.381	pregnancy
-0.236	os	0.380	year
-0.235	protein	0.378	mh-studies
-0.231	mice	0.372	mh-female
-0.231	serum	0.359	effect
-0.227	values	0.355	95
-0.223	model	0.354	age
-0.223	mm	0.349	mh-male

What's up with these??

One theory: these capture *numerical (post-hoc) analyses of data*, not primary studies

Features that predict media coverage

negative		*positive*	
-1.102	patients	0.617	exercise
-0.507	clinical	0.610	mh-data
-0.457	2012	0.604	mh-numerical mh-data
-0.393	survival	0.586	mh-numerical
-0.364	therapy	0.536	intake
-0.337	complications	0.531	mh-adult
-0.323	surgical	0.508	cancer
-0.321	response	0.500	mh-effects
-0.308	plasma	0.492	years
-0.300	pediatric	0.461	mh-child
-0.285	diagnostic	0.459	virus
-0.281	imaging	0.455	mh-aged
-0.275	2013	0.443	smoking
-0.267	management	0.442	influenza
-0.265	expression	0.428	mh-female mh-humans
-0.258	factors	0.418	consumption
-0.252	outcomes	0.407	incident
-0.250	score	0.407	women
-0.248	range	0.407	weight
-0.246	treatment	0.391	mh-humans
-0.246	function	0.389	exposure
-0.245	diabetes	0.383	asd
-0.242	review	0.381	pregnancy
-0.236	os	0.380	year
-0.235	protein	0.378	mh-studies
-0.231	mice	0.372	mh-female
-0.231	serum	0.359	effect
-0.227	values	0.355	95
-0.223	model	0.354	age
-0.223	mm	0.349	mh-male

What's up with these??

One theory: these capture *numerical (post-hoc) analyses of data*, not primary studies

This would also explain why *patients* and *clinical* **negatively** correlate with coverage

Additional evidence for this hypothesis

In an accompanying editorial in the journal, Ben Goldacre, author of the book *Bad Science*, noted that bad news tends to generate more coverage than good, and that less rigorous observational studies tend to generate more coverage than robust clinical trials, probably due to the applicability of the subject matter to lay readers.

On the predictive value of ‘000’

- The token ‘000’ ranked relatively high in terms of predicting coverage
  - This seemed weird

- It may be a stylistic thing -- real examples:
  - "Having more than 2 dermatologists per 100 000"
  - "Pediatric sudden cardiac death (SCD) occurs in an estimated 0.8 to 6.2 per 100 000 children annually"
  - "Prevalence per 10 000 population..”

- Wild conjecture: could expressing probabilities this way influence the likelihood that a journalist covers an article?

Press releases

- Journals often issue press releases; many journalists presumably follow these to decide what to cover

- We have so far ignored such factors (we do not know which articles in our Reuters corpus received press releases)

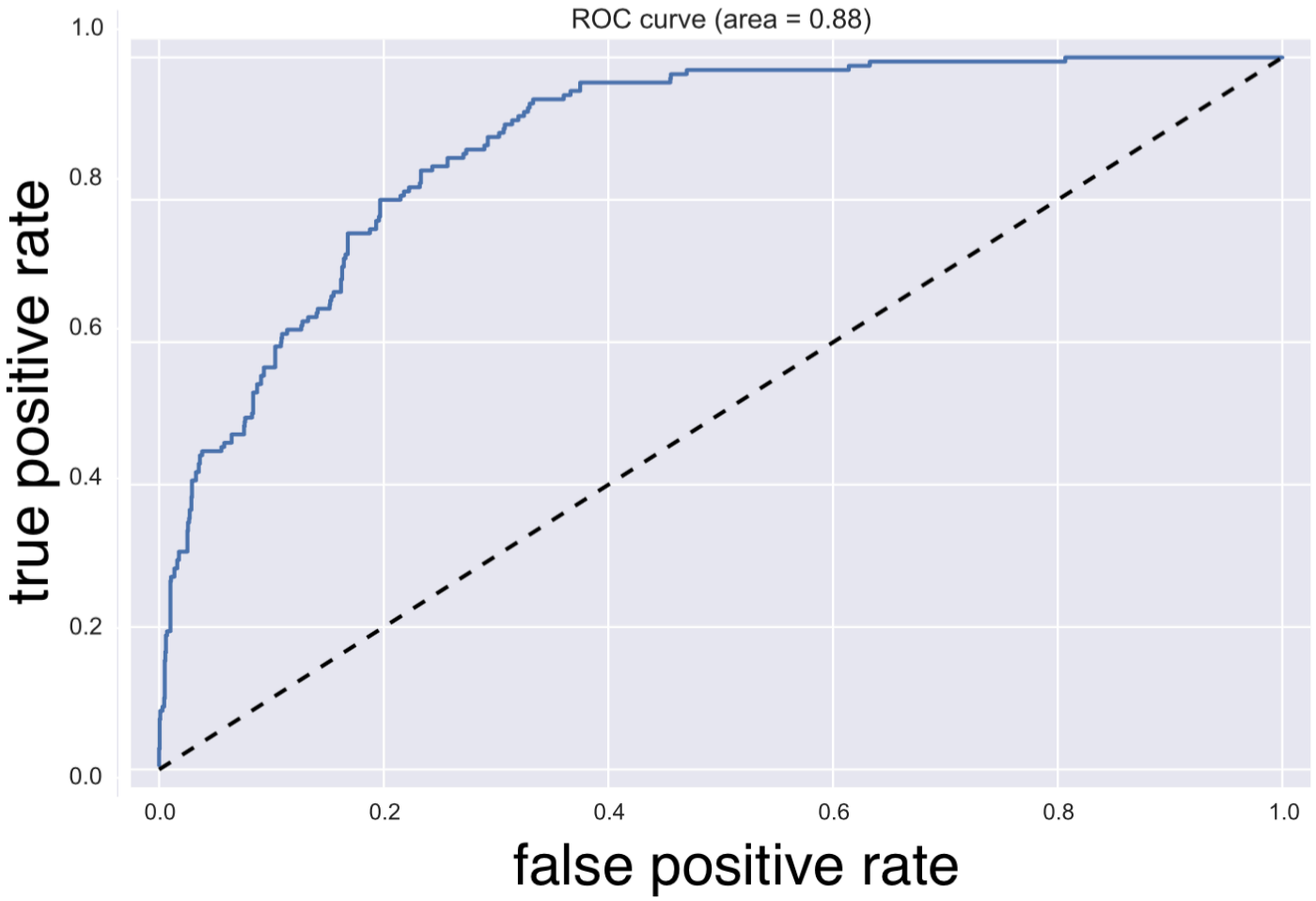
When do journals issue press releases?

- To answer this, we constructed a second corpus from all press releases issued by JAMA in 2013 and 2014

- In total: 1133 ‘positive’ articles for which press releases were issued
  - We again sampled ‘matched’ negative articles, which appeared in JAMA in the same year but were not given a press release

- Same learning setup as before

Can we predict press releases?



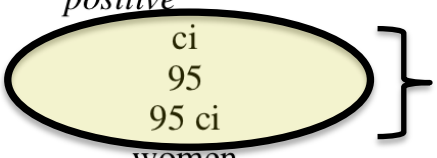
10-fold cross validation: average AUC 0.88, range (0.85, 0.92)

Features that predict press releases

negative		*positive*	
-0.615	patients	0.852	ci
-0.473	clinical	0.838	95
-0.357	dosing	0.808	95 ci
-0.356	sbp	0.750	women
-0.349	evidence	0.476	cancer
-0.318	injury	0.447	increased
-0.312	ezetimibe	0.433	mh-numerical
-0.310	functional	0.430	breast
-0.304	management	0.429	years
-0.302	review	0.425	mh-data
-0.294	patient	0.410	vs
-0.293	handover	0.407	mh-numerical mh-data
-0.290	schizophrenia	0.404	prevalence
-0.287	resection	0.373	men
-0.286	information	0.366	states
-0.281	mechanical	0.341	pregnancy
-0.277	aortic	0.341	insurance
-0.276	days	0.336	person-years
-0.274	hospitalization	0.325	tobacco
-0.274	acupuncture	0.319	rates
-0.273	score	0.316	breast cancer
-0.257	scores	0.311	maternal
-0.252	faculty	0.306	costs
-0.251	relapse	0.305	health
-0.245	bacteremia	0.303	chd
-0.244	gastric	0.303	cvd
-0.242	studies	0.283	smoking
-0.242	hcv	0.277	drinking
-0.239	continuity	0.266	child
-0.238	brain	0.262	age

Features that predict press releases

negative		*positive*	
-0.615	patients	0.852	ci
-0.473	clinical	0.838	95
-0.357	dosing	0.808	95 ci
-0.356	sbp	0.750	women
-0.349	evidence	0.476	cancer
-0.318	injury	0.447	increased
-0.312	ezetimibe	0.433	mh-numerical
-0.310	functional	0.430	breast
-0.304	management	0.429	years
-0.302	review	0.425	mh-data
-0.294	patient	0.410	vs
-0.293	handover	0.407	mh-numerical mh-data
-0.290	schizophrenia	0.404	prevalence
-0.287	resection	0.373	men
-0.286	information	0.366	states
-0.281	mechanical	0.341	pregnancy
-0.277	aortic	0.341	insurance
-0.276	days	0.336	person-years
-0.274	hospitalization	0.325	tobacco
-0.274	acupuncture	0.319	rates
-0.273	score	0.316	breast cancer
-0.257	scores	0.311	maternal
-0.252	faculty	0.306	costs
-0.251	relapse	0.305	health
-0.245	bacteremia	0.303	chd
-0.244	gastric	0.303	cvd
-0.242	studies	0.283	smoking
-0.242	hcv	0.277	drinking
-0.239	continuity	0.266	child
-0.238	brain	0.262	age



Note the emphasis on the p-value!

press releases

negative		*positive*	
-0.615	patients	0.852	ci
-0.473	clinical	0.838	95
-0.357	dosing	0.808	95 ci
-0.356	sbp	0.750	women
-0.349	evidence	0.476	cancer
-0.318	injury	0.447	increased
-0.312	ezetimibe	0.433	mh-numerical
-0.310	functional	0.430	breast
-0.304	management	0.429	years
-0.302	review	0.425	mh-data
-0.294	patient	0.410	vs
-0.293	handover	0.407	mh-numerical mh-data
-0.290	schizophrenia	0.404	prevalence
-0.287	resection	0.373	men
-0.286	information	0.366	states
-0.281	mechanical	0.341	pregnancy
-0.277	aortic	0.341	insurance
-0.276	days	0.336	person-years
-0.274	hospitalization	0.325	tobacco
-0.274	acupuncture	0.319	rates
-0.273	score	0.316	breast cancer
-0.257	scores	0.311	maternal
-0.252	faculty	0.306	costs
-0.251	relapse	0.305	health
-0.245	bacteremia	0.303	chd
-0.244	gastric	0.303	cvd
-0.242	studies	0.283	smoking
-0.242	hcv	0.277	drinking
-0.239	continuity	0.266	child
-0.238	brain	0.262	age

media coverage

negative		*positive*	
-1.102	patients	0.617	exercise
-0.507	clinical	0.610	mh-data
-0.457	2012	0.604	mh-numerical mh-data
-0.393	survival	0.586	mh-numerical
-0.364	therapy	0.536	intake
-0.337	complications	0.531	mh-adult
-0.323	surgical	0.508	cancer
-0.321	response	0.500	mh-effects
-0.308	plasma	0.492	years
-0.300	pediatric	0.461	mh-child
-0.285	diagnostic	0.459	virus
-0.281	imaging	0.455	mh-aged
-0.275	2013	0.443	smoking
-0.267	management	0.442	influenza
-0.265	expression	0.428	mh-female mh-humans
-0.258	factors	0.418	consumption
-0.252	outcomes	0.407	incident
-0.250	score	0.407	women
-0.248	range	0.407	weight
-0.246	treatment	0.391	mh-humans
-0.246	function	0.389	exposure
-0.245	diabetes	0.383	asd
-0.242	review	0.381	pregnancy
-0.236	os	0.380	year
-0.235	protein	0.378	mh-studies
-0.231	mice	0.372	mh-female
-0.231	serum	0.359	effect
-0.227	values	0.355	95
-0.223	model	0.354	age
-0.223	mm	0.349	mh-male

Moving forward

- **Improve matched sampling** by matching articles in the *same volume*, not only journal and year

- **Improve learning approach** train discriminative model on pairwise representation of positive/matched negative articles, instead of grouping these together into one training set
  - i.e., $y_i = \text{sign}\{\mathbf{w} \bullet (\mathbf{x}_i - \mathbf{x}_i^{\text{matched}})\}$ for all articles match matched for $\mathbf{x}_i$

Moving forward

- A new dataset (Sumners *et al.*) is available that includes articles followed from press releases to media coverage (or not): we will explore this

- Includes data on whether claims have been exaggerated; a potential future line of work might explore modeling this

Questions?

Data @ https://github.com/bwallace/w3phi-2015

byron.wallace@utexas.edu
http://byron.ischool.utexas.edu

mpaul39@gmail.com
http://cs.jhu.edu/~mpaul/

noemie.elhadad@columbia.edu
http://people.dbmi.columbia.edu/noemie/