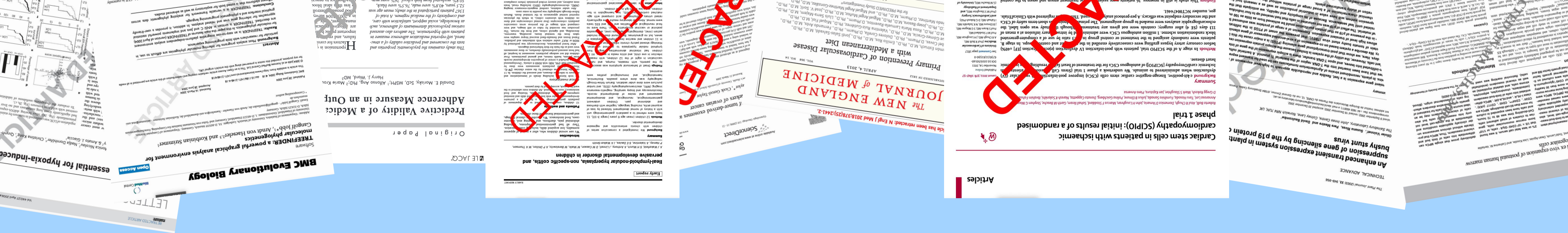
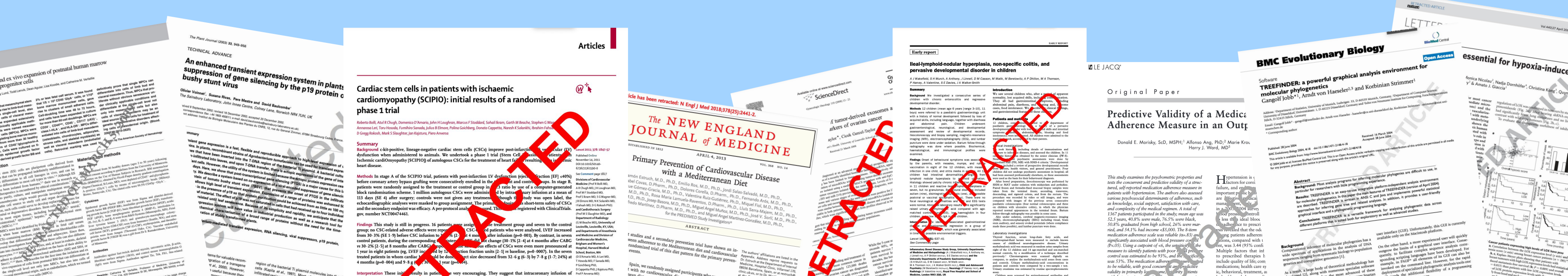


Predicting Paper Retractions

William Davis and Jack Kendrick



Motivation

- In theory, articles published in peer reviewed journals should hold up to scrutiny. When reading a paper, it is often taken for granted that the research is credible and trustworthy.
- However, sometimes bad papers fall through the cracks and end up being retracted.
- Retractions indicate seriously flawed and unreliable research, errors, fraud, ethical issues, or other serious concerns

vivo expansion of postnatal human marrow
nitor cells

The Plant Journal (2003) 33, 949–956

TECHNICAL ADVANCE

An enhanced transient expression system in plants
suppression of gene silencing by the p19 protein
bushy stunt virus

Olivier Voinnet, Susana Rivas, Pere Mestre and David Baulcombe

The Sainsbury Laboratory, John Innes Centre, Colney Lane, Norwich NR4 7UH, UK

Articles

Cardiac stem cells in patients with ischaemic cardiomyopathy (SCIPIO): initial results of a randomised phase 1 trial

Roberto Bolli, Atul R Chugh, Domenico D'Amario, John H Loughran, Marcus F Stoddard, Sohail Ikram, Garth M Beache, Stephen G Wag
Annarosa Leri, Toru Hosoda, Fumihiro Sanada, Julius B Elmore, Polina Goichberg, Donato Cappetta, Naresh K Solankhi, Ibrahim Fahsa
D Gregg Rokosh, Mark S Slaughter, Jan Kajstura, Piero Anversa

icle has been retracted: N Engl J Med 2018;378(25):2441-2.

The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812

APRIL 4, 2013

Available online at www.sciencedirect.com

ScienceDirect

f tumor-derived exosomes a
arkers of ovarian cancer

aylor, Cicek Gercel-Taylor

EARLY REPORT

Early report

Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children

A J Wakefield, S H Murch, A Anthony, J Linnell, D M Casson, M Malik, M Berelowitz, A P Dhillon, M A Thomson, P Harvey, A Valentine, S E Davies, J A Walker-Smith

LE JACQ

Original Paper

Predictive Validity of a Medica
Adherence Measure in an Outp

Donald E. Morisky, ScD, MSPH;[1] Alfonso Ang, PhD;[2] Marie Krou
Harry J. Ward, MD[5]

BMC Evolutionary Biology

Open Access

Software

TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics

Gangolf Jobb*[1], Arndt von Haeseler[2,3] and Korbinian Strimmer[1]

nature

RETRACTED ARTICLE

LETTER

Vol 440|27 April 2006

essential for hypoxia-induced

Monica Nicolau, Nadja Dornhöfer, Christina Kong, Quynh-Thu
& Amato J. Giaccia

Can we identify papers at a high risk of retraction?

The Plant Journal (2003) 33, 949–956
TECHNICAL ADVANCE
An enhanced transient expression system in plants based on suppression of gene silencing by the p19 protein of tomato bushy stunt virus
Articles
Cardiac stem cells in patients with ischaemic cardiomyopathy (SCIPIO): initial results of a randomised phase 1 trial
The NEW ENGLAND JOURNAL of MEDICINE
ScienceDirect
Early report
Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children
Original Paper
Predictive Validity of a Medica
Adherence Measure in an Outp
BMC Evolutionary Biology
TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics
essential for hypoxia-induced
RETRACTED ARTICLE
LETTER

*Data From RetractionWatch

Starting Small: PLOS One

- We concentrated on the journal PLOS One
- Relatively well known and respected, but has a high number of retractions

Rank	Journal	Retractions*
1	2011 International Conference on E-Business and E-Government	1280
2	2011 5th International Conference on Bioinformatics and Biomedical Engineering	1084
3	PLoS One	944
4	Journal of Physics: Conference Series	878

The Dataset

- ➔ We collected data from the OpenAlex database using the PyAlex API
- ➔ ***Huge*** amounts of data available (including retractions)
- ➔ We used the raw data from OpenAlex to build features of interest
  - ◆ Fraction of authors that have previous retractions, whether any authors come from institutions with many retractions, etc.
- ➔ Our dataset consists of papers published in 2010-2020: 424,223 papers, with 797 retractions

Challenge: Retracted vs Non-Retracted Classes are massively unbalanced

An enhanced transient expression system in plants based on suppression of gene silencing by the p19 protein of tomato bushy stunt virus
Cardiac stem cells in patients with ischaemic cardiomyopathy (SCIPIO): initial results of a randomised phase 1 trial
Articles
The NEW ENGLAND JOURNAL of MEDICINE
Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children
Early report
Predictive Validity of a Medica Adherence Measure in an Outp
Original Paper
BMC Evolutionary Biology
TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics
essential for hypoxia-induced
RETRACTED ARTICLE

Our Approach

- → Baseline logistic regression model
- → Compare more baseline to nearest neighbour, random forest, and SVC classification methods.

Key Performance Indicators: F_1-score, Precision, and Recall

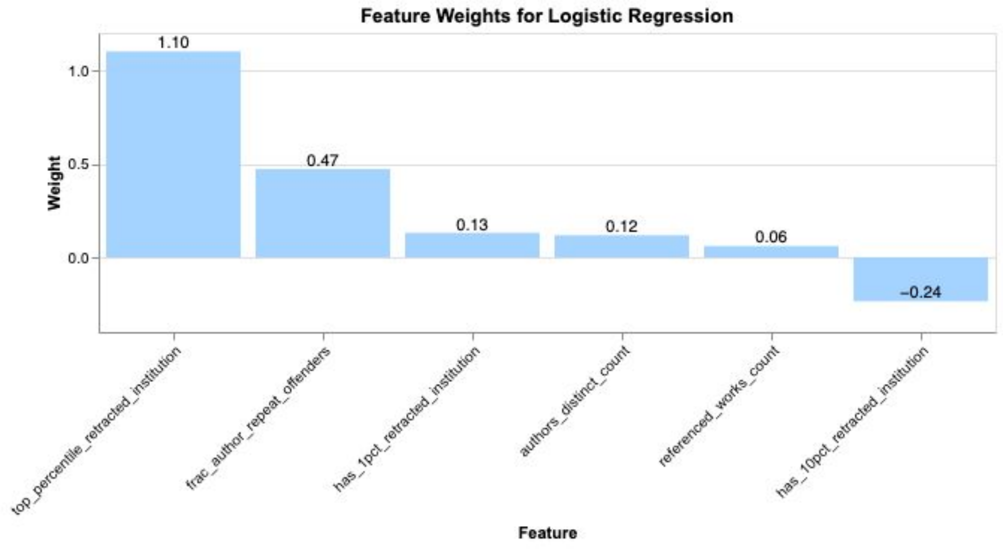
Baseline Model

- → Used **forward stepwise subset selection** to choose features.
- → The most informative feature is the proportion of authors on a paper that have been previously retracted
- → Second most informative feature is a measure of how many retractions any institution associated to the paper has received.
- → Results of subset selection correspond to our intuition on features that correlate to a risk of retraction

An enhanced transient expression system in plants based on suppression of gene silencing by the p19 protein of tomato bushy stunt virus
Cardiac stem cells in patients with ischaemic cardiomyopathy (SCIPIO): initial results of a randomised phase 1 trial
The NEW ENGLAND JOURNAL of MEDICINE
Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children
Predictive Validity of a Medica Adherence Measure in an Outp
BMC Evolutionary Biology
TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics
essential for hypoxia-induced

Baseline Model

F_1-Score:	0.169
Precision:	0.394
Recall:	0.107



TECHNICAL ADVANCE

An enhanced transient expression system in plants based on suppression of gene silencing by the p19 protein o

Articles

Cardiac stem cells in patients with ischaemic cardiomyopathy (SCIPIO): initial results of a randomised phase 1 trial

The NEW ENGLAND JOURNAL of MEDICINE

Early report

Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children

Original Paper

Predictive Validity of a Medica Adherence Measure in an Outp

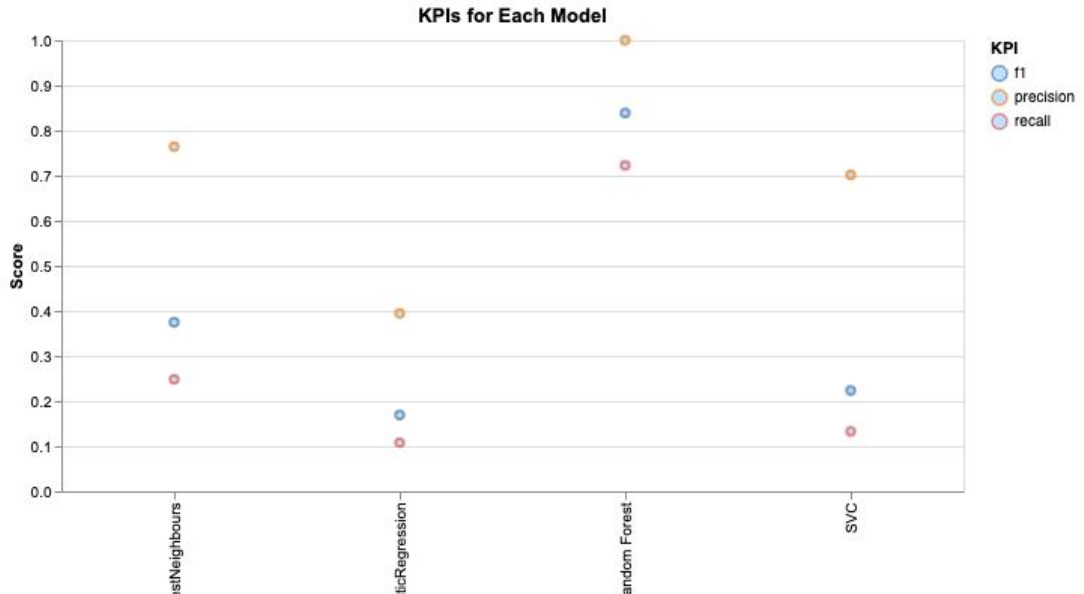
BMC Evolutionary Biology

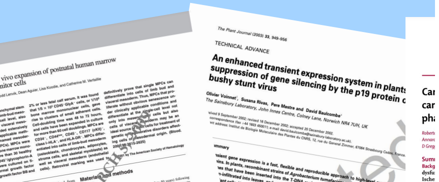
TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics

essential for hypoxia-induced

Model Selection

- → Tested k-Nearest Neighbors (k = 1, 2,…., 10), random forests, and support vector classification.
- → We used stratified 10-fold cross-validation to choose hyperparameters.





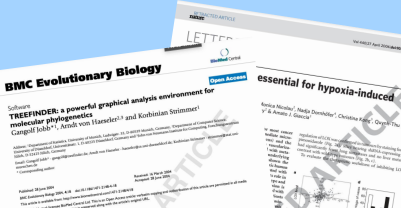
vivo expansion of postnatal human marrow
itor cells

The Plant Journal (2003) 33, 949–956

TECHNICAL ADVANCE

An enhanced transient expression system in plants
suppression of gene silencing by the p19 protein o
bushy stunt virus

Olivier Voinnet, Susana Rivas, Pere Mestre and David Baulcombe



BMC Evolutionary Biology

Open Access

Software

TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics

Gangolf Jobb, Arndt von Haeseler and Korbinian Strimmer

essential for hypoxia-induced

Results

→ Random Forest classification far outperformed our other models in training.

Final Model: Random Forest with 500 estimators, max depth of 20

F_1 :	***0.289***
Precision:	***0.691***
Recall:	***0.182***
Accuracy:	***0.998***

vivo expansion of postnatal human marrow
nitor cells

The Plant Journal (2003) 33, 949–956

TECHNICAL ADVANCE

An enhanced transient expression system in plants
suppression of gene silencing by the p19 protein o
bushy stunt virus

Olivier Voinnet, Susana Rivas, Pere Mestre and David Baulcombe

Articles

Cardiac stem cells in patients with ischaemic cardiomyopathy (SCIPIO): initial results of a randomised phase 1 trial

Roberto Bolli, Atul R Chugh, Domenico D'Amario, John H Loughran, Marcus F Stoddard, Sohail Ikram, Garth M Beache, Stephen G Wag
Annarosa Leri, Toru Hosoda, Fumihiro Sanada, Julius B Elmore, Polina Goichberg, Donato Cappetta, Naresh K Solankhi, Ibrahim Fahsa
D Gregg Rokosh, Mark S Slaughter, Jan Kajstura, Piero Anversa

icle has been retracted: N Engl J Med 2018;378(25):2441-2.

The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 APRIL 4, 2013

ScienceDirect

f tumor-derived exosomes a
arkers of ovarian cancer

EARLY REPORT

Early report

Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children

A J Wakefield, S H Murch, A Anthony, J Linnell, D M Casson, M Malik, M Berelowitz, A P Dhillon, M A Thomson, P Harvey, A Valentine, S E Davies, J A Walker-Smith

LE JACQ

Original Paper

Predictive Validity of a Medica
Adherence Measure in an Outp

Donald E. Morisky, ScD, MSPH; Alfonso Ang, PhD; Marie Krou
Harry J. Ward, MD

BMC Evolutionary Biology

Open Access

Software

TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics

Gangolf Jobb, Arndt von Haeseler and Korbinian Strimmer

RETRACTED ARTICLE

nature

LETTE

essential for hypoxia-induced

Future Directions

- → Expand dataset to include other journals and features relating to funding sources, journal publishers, etc.
- → Investigate papers published in the post ChatGPT-era. Can we use ‘Contains AI generated text’ as a feature?

vivo expansion of postnatal human marrow itor cells

The Plant Journal (2003) 33, 949–956

TECHNICAL ADVANCE

An enhanced transient expression system in plants based on suppression of gene silencing by the p19 protein of tomato bushy stunt virus

Olivier Voinnet, Susana Rivas, Pere Mestre and David Baulcombe

Articles

Cardiac stem cells in patients with ischaemic cardiomyopathy (SCIPIO): initial results of a randomised phase 1 trial

Roberto Bolli, Atul R Chugh, Domenico D'Amario, John H Loughran, Marcus F Stoddard, Sohail Ikram, Garth M Beache, Stephen G Wagner, Annarosa Leri, Toru Hosoda, Fumihiro Sanada, Julius B Elmore, Polina Goichberg, Donato Cappetta, Naresh K Solankhi, Ibrahim Fahsah, D Gregg Rokosh, Mark S Slaughter, Jan Kajstura, Piero Anversa

This article has been retracted: N Engl J Med 2018;378(25):2441-2.

The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812

APRIL 4, 2013

Available online at www.sciencedirect.com

ScienceDirect

of tumor-derived exosomes as markers of ovarian cancer

EARLY REPORT

Early report

Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children

A J Wakefield, S H Murch, A Anthony, J Linnell, D M Casson, M Malik, M Berelowitz, A P Dhillon, M A Thomson, P Harvey, A Valentine, S E Davies, J A Walker-Smith

LE JACQ

Original Paper

Predictive Validity of a Medication Adherence Measure in an Outpatient Setting

Donald E. Morisky, ScD, MSPH; Alfonso Ang, PhD; Marie Krousel-Wood, Harry J. Ward, MD

BMC Evolutionary Biology

Open Access

Software

TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics

Gangolf Jobb, Arndt von Haeseler and Korbinian Strimmer

RETRACTED ARTICLE

nature

LETTERS

Vol 440 | 27 April 2006

essential for hypoxia-induced

Acknowledgements

- ➔ Many thanks to our mentor Greg Edwards for his help and expertise throughout this project
- ➔ Thank you to the Erdos Institute for providing us with this excellent bootcamp.

An enhanced transient expression system in plants based on suppression of gene silencing by the p19 protein of tomato bushy stunt virus

Articles

Cardiac stem cells in patients with ischaemic cardiomyopathy (SCIPIO): initial results of a randomised phase 1 trial

The NEW ENGLAND JOURNAL of MEDICINE

Early report

Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children

Original Paper

Predictive Validity of a Medica Adherence Measure in an Outp

BMC Evolutionary Biology

TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics

essential for hypoxia-induced