Reproducibility and Cognitive Issues in Publications Based on Big Data

Želimir Kurtanjek

University of Zagreb

Faculty of Food Technology and Biotechnology

* retired

Outline

- Big Data critical issues
  - Life sciences, technical sciences, social sciences,
- Prominent examples
- Sources of contradictions
- Data forensics
- Causality, model validation and p-value inference
- Propositions of editorial corrective measures
- Conclusions

How big are „Big Data" and its two faces

- Data size (EU human genome project)
- 3 x10^9 (base pairs) x 10^7 human x 10^3 phenotypes = 10^19 numerical data
- „Gold bars" and „new oil" versus „card castles"

Big Data are omnipresent

- Life sciences: Mendelian large cohort studies, genetics, proteomics, glycomics, metabolomics, nutrigenomics..
- Technical sciences: AI, G5, Internet of Things, Robotics
- Social sciences: behavioral studies, social networks, ..
- Economy: Financial engineering, marketing, managment
- Government: e-government policies, cyber security ..

Big data with „two faces”

➢Big data have high market value and are power engine („new oil”) of G5 economy

➢Big data research produces „houses of cards”, i.e. look plausible (nice) but do not „touch”

What are problems with Big Data research publications ?

Retraction Watch Database User Guide

Welcome to our database. We've prepared this document to help you get started, and to answer some questions that are likely to come up. This document will evolve as users have more questions, so please feel free to contact us at team@retraction-watch.com.

Top 10 most high impact retracted papers are in field of Life Science

Examples

Article	Year of retraction	Citing Articles before re-traction	Citing Articles after re-traction	Total cites (journals indexed by Web of Science)
1. Primary Prevention of Cardiovascular Disease with a Mediterranean Diet. N Engl J Med April 4, 2013	2018	1879	271	2150

?????

Version: 1.0.5.5

The Retraction Watch Database

Please see this user guide before you get started

Author(s): Type to search

Country(s):

Title: Primary Prevention of Cardiovascular Disease with a Mediterranean Diet

Reason(s) for Retraction:

Subject(s):

Article Type(s):

Journal:

Publisher:

Affiliation(s):

Notes:

URL:

Clear Search

Search

Retraction or Other Notices Title/Subject(s)/Journal --- Publisher/Affiliation(s)/Retraction Watch Post URL(s) 1 Item(s) Found	Reason(s)	Author(s)
Primary Prevention of Cardiovascular Disease with a Mediterranean Diet (HSC) Medicine - Cardiology; (HSC) Medicine - Cardiovascular; (HSC) Nutrition; (HSC) Public Health and Safety; *The New England Journal of Medicine* --- ***Massachusetts Medical Society*** Centro de Investigacion Biomedica en Red de Fisiopatologia de la Obesidad y and the PREDIMED (Prevención con Dieta Mediterránea) Network (RD 06/0045), Instituto de Salud Carlos III, Madrid Department of Internal Medicine and Lipid Clinic, Department of Endocrinology and Nutrition, Institut d'Investigacions Biomediques August Pi I Sunyer, Hospital Clinic, University of Barcelona, Barcelona Human Nutrition Department, Hospital Universitari Sant Joan, Institut d'Investigacio Sanitaria Pere Virgili, Universitat Rovira i Virgili, Reus Cardiovascular and Nutrition Research Group, Institut de Recerca Hospital del Mar, Barcelona Department of Preventive Medicine, University of Valencia, Valencia Department of Cardiology, University Hospital of Alava, Vitoria Department of Preventive Medicine, University of Malaga, Malaga Instituto de la Grasa, Consejo Superior de Investigaciones Cientificas, Seville Institute of Health Sciences (IUNICS), University of Balearic Islands, and Hospital Son Espases, Palma de Mallorca Department of Family Medicine, Primary Care Division of Seville, San Pablo Health Center, Seville Department of Nutrition and Food Science, School of Pharmacy, Xarxa de Referencia en Tecnologia dels Aliments, Instituto de Investigacion en	+Error in Analyses +Error in Methods +Error in Results and/or Conclusions +Retract and Replace	Ramon Estruch Emilio Ros Jordi Salas-Salvado Maria Isabel Covas Dolores Corella Fernando Aros Enrique Gomez-Gracia Valentina Ruiz-Gutierrez Miguel Fiol Jose Lapetra Rosa Maria Lamuela-Raventos Lluis Serra-Majem Xavier Pinto Joseph Basora Miguel Angel Munoz Jose V Sorli

Scholarly articles for **Potti genomic signature**

Genomic signatures to guide the use of ... - **Potti** Cited by 620

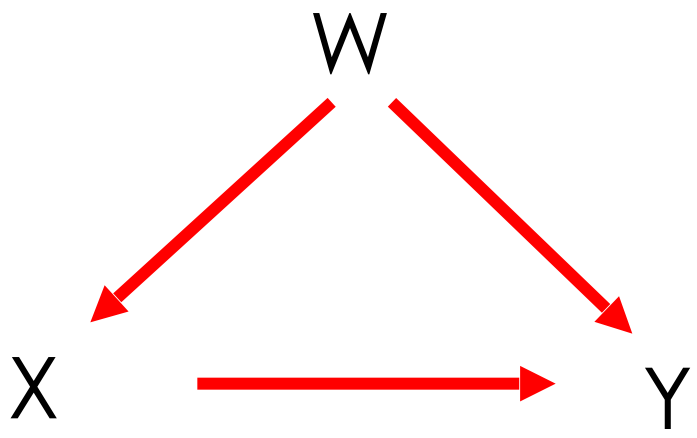
Genomic strategies for personalized cancer therapy - Garman - Cited by 62

Mining gene expression profiles: expression **signatures** ... - Nevins - Cited by 180

The Retraction Watch Database
Please see this user guide before you get started

Genomic signatures to guide the use of chemotherapeutics (BLS) Biochemistry; (BLS) Biology - Cancer; (BLS) Biology - Cellular; (BLS) Genetics; (HSC) Medicine - Drug Design; (HSC) Medicine - Oncology; (HSC) Medicine - Pharmacology; *Nature Medicine* — *Springer - Nature Publishing Group*	+Investigation by Company/Institution +Investigation by Third Party +Results Not Reproducible	Anil Potti Holly K Dressman Andrea Bild Richard F Riedel Gina Chan Robyn Sayer Janiel Cragun Hope Cottrill Michael J Kelley Rebecca Petersen David Harpole Jeffrey Marks Andrew Berchuck Geoffrey S Ginsburg Phillip G Febbo Johnathan Lancaster Joseph R Nevins
Duke Institute for Genome Sciences and Policy, Duke University, Durham, North Carolina Department of Medicine, Duke University Medical Center, Durham, North Carolina Department of Molecular Genetics and Microbiology, Duke University Medical Center, Durham, North Carolina Division of Gynecologic Surgical Oncology, H. Lee Moffitt Cancer Center & Research Institute, University of South Florida, Tampa, Florida Department of Surgery, Duke University Medical Center, Durham, North Carolina Department of Obstetrics and Gynecology, Duke University Medical Center, Durham, North Carolina		
http://retractionwatch.com/2011/01/07/nature-medicine-makes-it-official-retracting-anil-potti-paper/ http://retractionwatch.com/2010/11/19/another-update-on-anil-potti-co-author-asks-nature-medicine-to-retract-paper/		

Causality structure of Big Data research



$Y=f(X)$	Causal relation
$Y=f(X, W\approx 0)$	Randomized trials
$Y=f(X, W)$	Adjusted confounders, Propensity score
$Y=f(X, W)$	Confounded causality

W confiders of high dimension, some unobserved

X causality X={0,1}

Y effect Y ={0,1}, Y= {R}

Causality analysis is study of effect of counterfactuals

Main problems with Big Data published research are due to:

- ➢Lack of causality model (structure)
- ➢Missing methodology for confounder adjustments
- ➢Unvalidated data (experimental procedures)
- ➢Unvalidated model predictions
- ➢Unreported confidence bounds for inference parameters (p-values)

The problems are of systemic, „deep" nature and require **main changes in journal editorial policies**

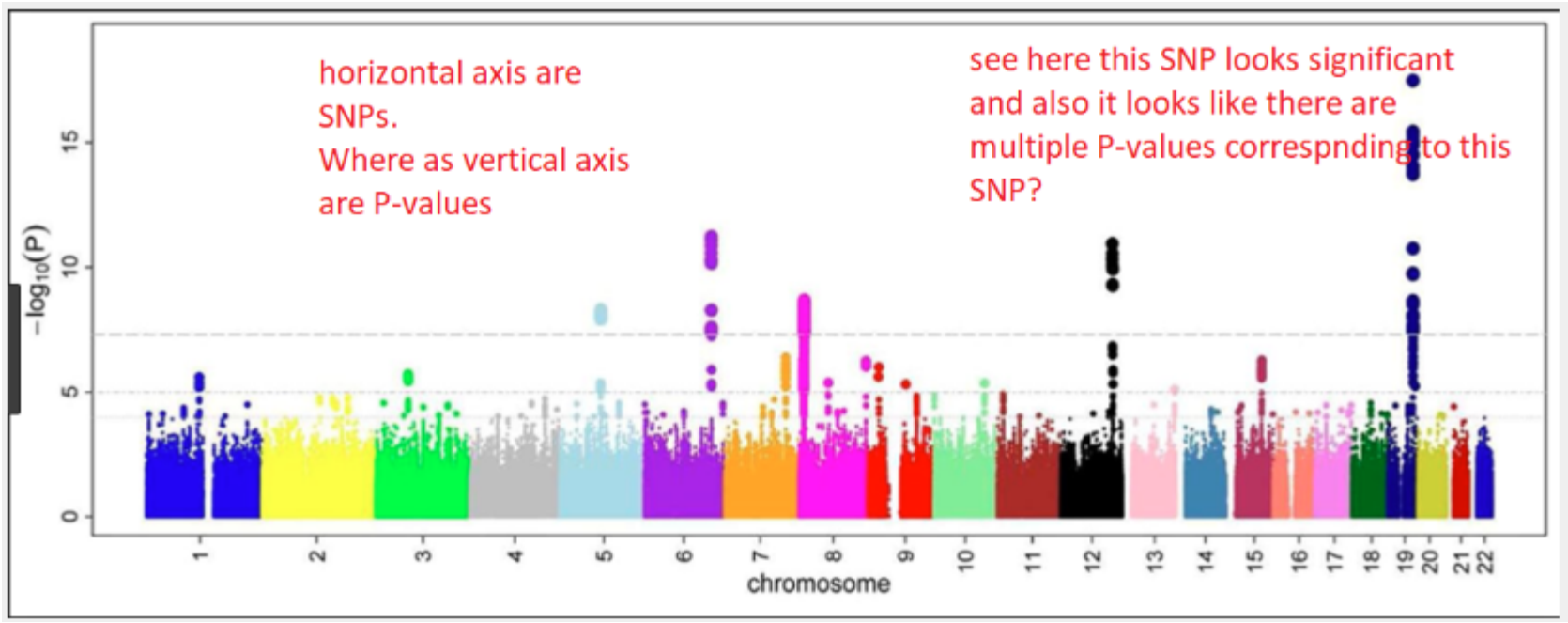
Software tools available to editorial boards (reviewers) for „check" of Big Data manuscripts

- Data forensics (Benford „law")
- Stat-checking software

COMMENT · 20 MARCH 2019

Scientists rise up against statistical significance

Valentin Amrhein, Sander Greenland, Blake McShane and more than 800 signatories call for an end to hyped claims and the dismissal of possibly crucial effects.



GWAS association

NATURE | TOOLBOX

Stat-checking software stirs up

Researchers debate whether using a program to automatically detect inconsistencies in papers improves the literature, or raises false alarms.

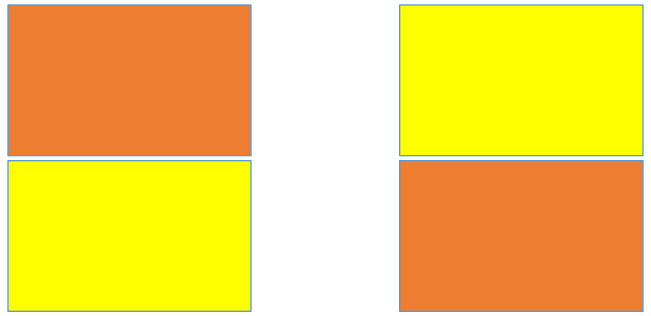
Monya Baker

Michèle Nuijten and her colleagues found rampant inconsistencies when they unleased statcheck on the psychological literature. The program scans articles for statistical results, redoes the calculations and checks that the numbers match. It went through 30,717 papers to identify 16,695 that tested hypotheses using statistics.

Basic methodologies for Big Data validation
(that should be imposed by editorial policies)

Model validation by →

Data set folding



Inference validation by →

Data set bootstrapping

Data forensics

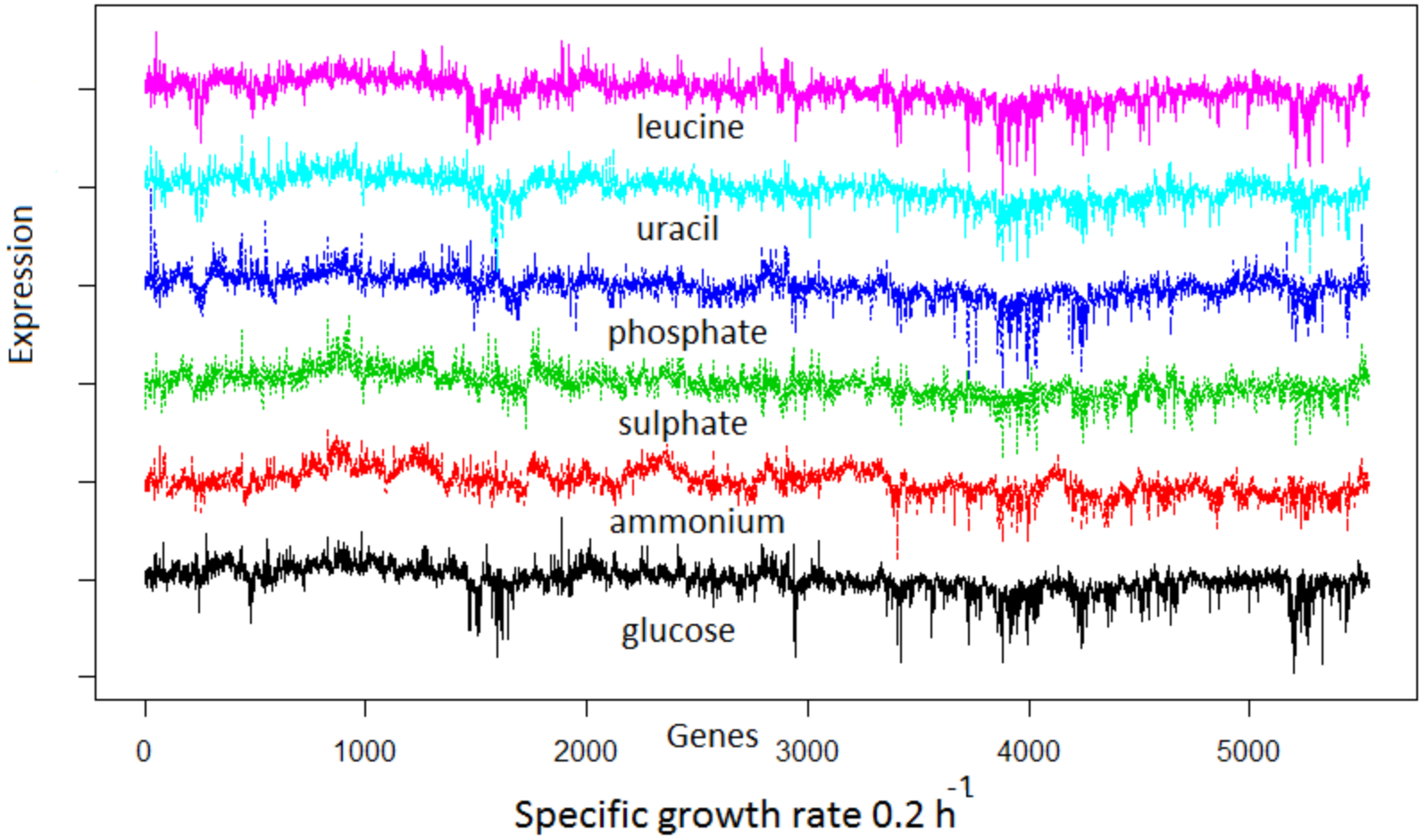
What is Benford's Law and why is it important for data science?

Benford's law tells us about expected distribution of significant digits in a diverse set of naturally occurring datasets and how this can be used for anomaly or fraud detection in scientific or technical publications !!!!

The first record on data sets from 1881

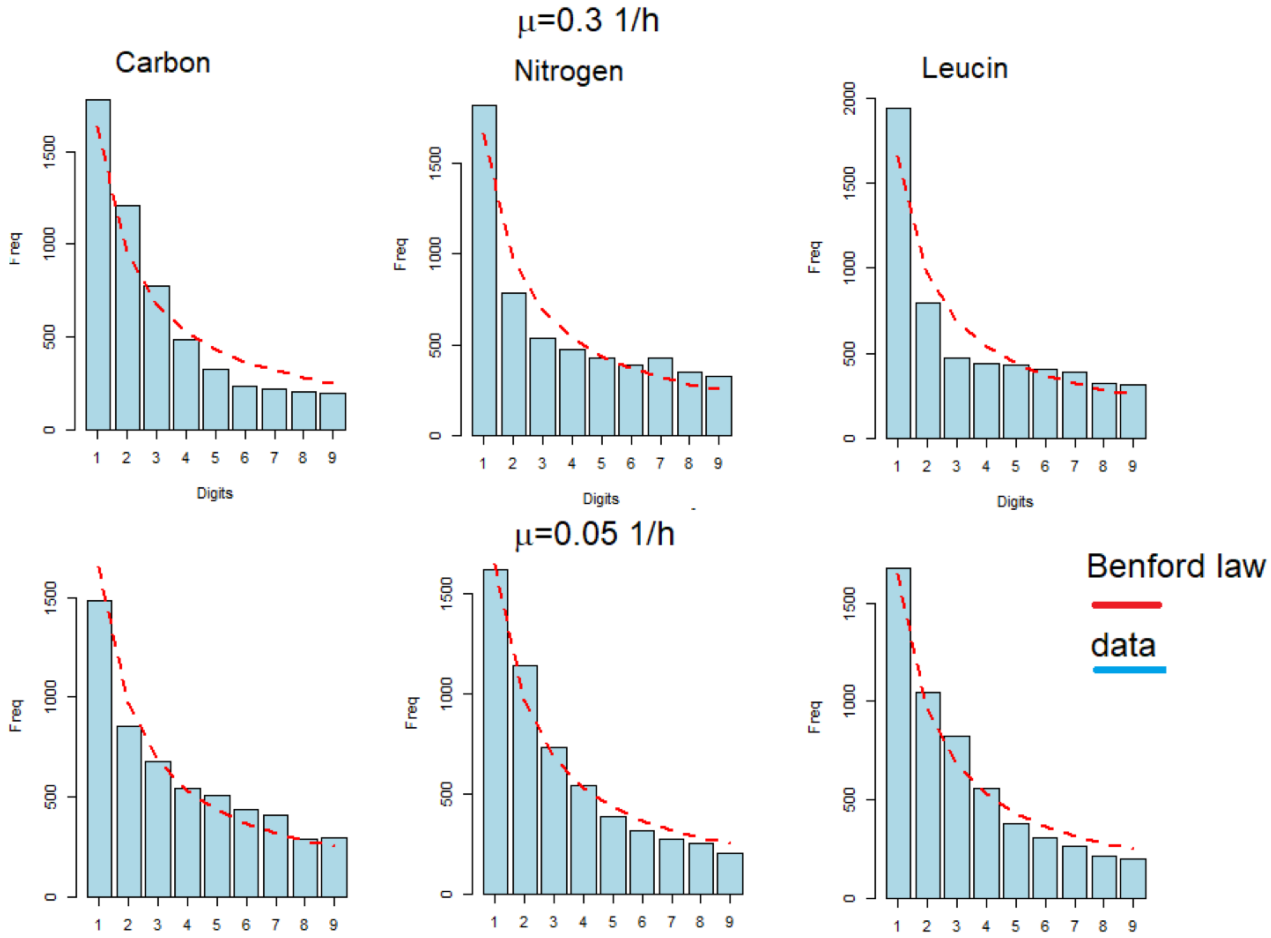
Mathematical proof published in 1996 in paper:
A Statistical Derivation of the Significant-Digit Law
Theodore P. Hill
School of Mathematics and Center for Applied Probability
Georgia Institute of Technology
Atlanta, GA

Yeast GW expression (mRNA) data



Data source: M. Brauer at al. http://growthrate.princeton.edu/
"https://4va.github.io/biodatasci/data/brauer2007_tidy.csv"

Yeast GW gene (mRNA) expressions under substrate limitations
Data forensics by Benford's „law"



Benford **law does not validate for N=2**, hence mRNA expression data error level is **~10 %**

Conclusions

➢Advances of high throughput experimental techniques and information technologies led to Big Data science a dominant trend in life sciences, also in other scientific fields (social, economy, production technologies, …)

➢ Due to new technologies, complexity and size of Big Data research for science publishers have resulted in pressure to change and adjust editorial policies to meet challenges of data validation and cognitive contribution of published manuscripts.

➢High impact factor of retracted (erroneous cognition) Big Data longitudinal research in human health fields makes them seriously damaging.

➢The „old policy" that a single reviewer is competent for a whole content of a submitted manuscript is mostly untrue. A group of experts in different aspects of Big Data projects should cooperate and produce a single integrated review („triangulation by reviewers").

➢ Policies of Open science data, publication and reviews is essential for research in life sciences.

➢ To editorial boards are available methodologies and software supports for validation of model predictions and cognitive inferences in Big Data research.

➢ Most of issues won't be solved with a single rule or policy, the best solution available is to just start discussing ways how we can improve practice of Big Data and related analytical fields.