What is health data science?

STA 198: Introduction to Health Data Science

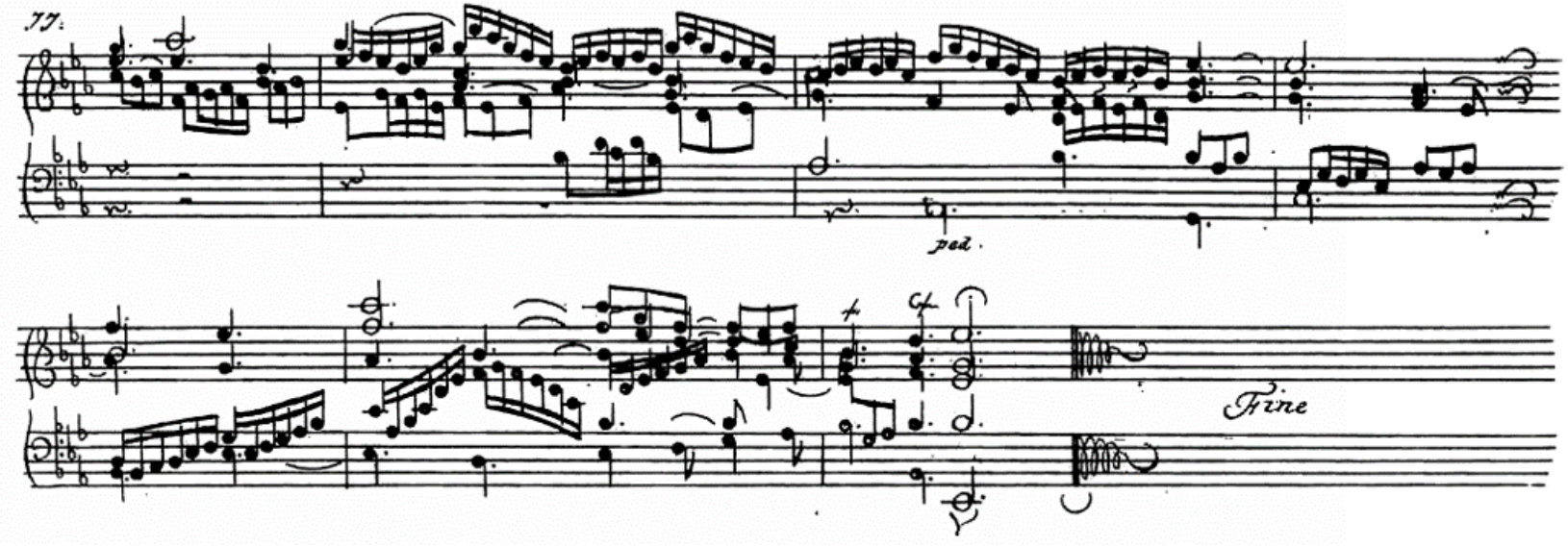
Yue Jiang

May 18, 2020

The following material was used by Yue Jiang during a live lecture.

Without the accompanying oral comments, the text is incomplete as a record of the presentation.

Who are you?



Who are you?

NIH NIEHS

Lilly

genzyme

putnam associates

INSIGHT. CLARITY. STRATEGY.

Who are you?

RESEARCH ARTICLE

Use of detailed family history data to improve risk prediction,with application to breast cancer screening

Yue Jiang, Clarice R. Weinberg, Dale P. Sandler, Shanshan Zhao

OBSERVATIONS AND RESEARCH

The Early Experience With Vedolizumab in the United States

Bharati Kochar, MD, MSCR,[*,†] Yue Jiang, BS,[‡] Aaron Winn, PhD,[§] Edward L. Barnes, MD, MPH,[*,†] Christopher F. Martin, MSPH,[†] Millie D. Long, MD, MPH,[*,†] and Michael D. Kappelman, MD, MPH[*,†]

Opioid prescriptions are associated with hepatic encephalopathy in a national cohort of patients with compensated cirrhosis

Andrew M. Moon, Yue Jiang, Shari S. Rogal, Elliot B. Tapper, Sarah R. Lieber, A. Sidney Barritt IV

ORIGINAL RESEA
RESEARCH SPONS

Decreasing Colectomy Rate for Ulcerative Colitis in the United States Between 2007 and 2016: A Time Trend Analysis

Edward L. Barnes, MD, MPH,[*,†,‡] Yue Jiang,[§] Michael D. Kappelman, MD, MPH,[†,‡,¶] Millie D. Long, MD, MPH,[*,†,‡] Robert S. Sandler, MD, MPH,[*,‡] Alan C. Kinlaw, PhD,[‖,**,a] and Hans H. Herfarth, MD, PhD[*,†,‡,a]

ACKD

Competing Risk of Death With End-Stage Renal Disease in Diabetic Kidney Disease

Yue Jiang, Jason P. Fine, and Amy K. Mottl

Clinical TRANSPLANTATION

The impact of post-transplant diabetes mellitus on liver transplant outcomes

Sarah R. Lieber[1] | Ruth-Ann Lee[2] | Yue Jiang[3] | Claire
Kristen Szempruch[2] | David A. Gerber[5] | Chirag S. Desai[5]
A. Sidney Barritt IV[1]

RESEARCH ARTICLE

In inpatients with cirrhosis opioid use is common and associated with length of stay and persistent use post-discharge

Andrew M. Moon, Yue Jiang, Shari S. Rogal, Jasper Becker, A. Sidney Barritt IV

What have I gotten myself into?

STA 198L is a rigorous ten-week online introduction to health data science that...

- ...provides a tour of basic statistical methods commonly encountered in public health and biomedical research,
- ...emphasizes understanding of the methods, using them to arrive at data-driven decisions, effective communication of such results, and critically assessing existing evidence,
- ...is motivated by timely, relevant examples from current research in biomedicine, epidemiology, and public health policy, and
- ...utilizes modern software such as RStudio and GitHub to reproducibly examine and manipulate data to make sound scientific conclusions.

What IS health data science?

A process that converts data into useful information, whereby practitioners

1. form a question of interest,
2. collect and summarize data,
3. and interpret the results

What is health data science good for?

The NEW ENGLAND JOURNAL of MEDICINE

SPECIAL ARTICLE

Mortality in Puerto Rico after Hurricane Maria

Nishant Kishore, M.P.H., Domingo Marqués, Ph.D., Ayesha Mahmud, Ph.D., Mathew V. Kiang, M.P.H., Irmary Rodriguez, B.A., Arlan Fuller, J.D., M.A., Peggy Ebner, B.A., Cecilia Sorensen, M.D., Fabio Racy, M.D., Jay Lemery, M.D., Leslie Maas, M.H.S., Jennifer Leaning, M.D., S.M.H., Rafael A. Irizarry, Ph.D., Satchit Balsari, M.D., M.P.H., and Caroline O. Buckee, D.Phil.

NEJM (July, 2018)

What is health data science good for?

ARTICLES
https://doi.org/10.1038/s41477-018-0263-1
nature plants

Decreases in global beer supply due to extreme drought and heat

Wei Xie[1*], Wei Xiong[2,3,4], Jie Pan[2], Tariq Ali[1], Qi Cui[5], Dabo Guan[6,7*], Jing Meng[8], Nathaniel D. Mueller[9], Erda Lin[2*] and Steven J. Davis[9,10]

Nature *Plants* (October, 2018)

What is health data science good for?

SCIENCE TRANSLATIONAL MEDICINE | RESEARCH ARTICLE

GENETIC DIAGNOSIS

Diagnosis of genetic diseases in seriously ill children by rapid whole-genome sequencing and automated phenotyping and interpretation

Michelle M. Clark[1], Amber Hildreth[1,2,3], Sergey Batalov[1], Yan Ding[1], Shimul Chowdhury[1], Kelly Watkins[1], Katarzyna Ellsworth[1], Brandon Camp[1], Cyrielle I. Kint[4], Calum Yacoubian[5], Lauge Farnaes[1,2], Matthew N. Bainbridge[1,6], Curtis Beebe[7], Joshua J. A. Braun[1], Margaret Bray[8], Jeanne Carroll[1,2], Julie A. Cakici[1], Sara A. Caylor[1], Christina Clarke[1], Mitchell P. Creed[9], Jennifer Friedman[1,10], Alison Frith[5], Richard Gain[5], Mary Gaughran[1], Shauna George[7], Sheldon Gilmer[7], Joseph Gleeson[1,10], Jeremy Gore[11], Haiying Grunenwald[12], Raymond L. Hovey[1], Marie L. Janes[1], Kejia Lin[7], Paul D. McDonagh[8], Kyle McBride[7], Patrick Mulrooney[1], Shareef Nahas[1], Daeheon Oh[1], Albert Oriol[7], Laura Puckett[1], Zia Rady[1], Martin G. Reese[13], Julie Ryu[1,2], Lisa Salz[1], Erica Sanford[1,2], Lawrence Stewart[7], Nathaly Sweeney[1,2], Mari Tokita[1], Luca Van Der Kraan[1], Sarah White[1], Kristen Wigby[1,2], Brett Williams[5], Terence Wong[1], Meredith S. Wright[1], Catherine Yamada[1], Peter Schols[4], John Reynders[8], Kevin Hall[12], David Dimmock[1], Narayanan Veeraraghavan[1], Thomas Defay[8], Stephen F. Kingsmore[1]*

Science *Translational Medicine* (April, 2019)

What is health data science good for?

ESMO GOOD SCIENCE BETTER MEDICINE BEST PRACTICE

Annals of Oncology 29: 1836–1842, 2018
doi:10.1093/annonc/mdy166
Published online 28 May 2018

ORIGINAL ARTICLE

Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists

H. A. Haenssle[1*,†], C. Fink[1†], R. Schneiderbauer[1], F. Toberer[1], T. Buhl[2], A. Blum[3], A. Kalloo[4], A. Ben Hadj Hassen[5], L. Thomas[6], A. Enk[1] & L. Uhlmann[7]

Annals of Oncology (August, 2018)

What is health data science good for?

Ramucirumab plus erlotinib in patients with untreated, *EGFR*-mutated, advanced non-small-cell lung cancer (RELAY): a randomised, double-blind, placebo-controlled, phase 3 trial

*Kazuhiko Nakagawa, Edward B Garon, Takashi Seto, Makoto Nishio, Santiago Ponce Aix, Luis Paz-Ares, Chao-Hua Chiu, Keunchil Park, Silvia Novello, Ernest Nadal, Fumio Imamura, Kiyotaka Yoh, Jin-Yuan Shih, Kwok Hung Au, Denis Moro-Sibilot, Sotaro Enatsu, Annamaria Zimmermann, Bente Frimodt-Moller, Carla Visseren-Grul, Martin Reck, for the RELAY Study Investigators**

Lancet *Oncology* (October, 2019)

What is health data science good for?

Research

JAMA Internal Medicine | Original Investigation

Comparison of Hospital Mortality and Readmission Rates for Medicare Patients Treated by Male vs Female Physicians

Yusuke Tsugawa, MD, MPH, PhD; Anupam B. Jena, MD, PhD; Jose F. Figueroa, MD, MPH; E. John Orav, PhD; Daniel M. Blumenthal, MD, MBA; Ashish K. Jha, MD, MPH

JAMA *Internal Medicine* (February, 2017)

Back to the statistical process

1. Forming a question of interest
2. Collecting and summarizing data
3. Interpreting the results

Identifying the population and question of interest

The population is the group we'd like to learn something about:

- What is the prevalence of diabetes among U.S. adults, and has it changed over time?
- Is there a relationship between tumor type and five-year mortality in breast cancer patients?
- Does the average amount of caffeine vary by vendor in 12 oz. cups of coffee at Duke coffee shops?

If we had data from every unit in the population, we could just calculate what we wanted and be done!

Sampling from the population

Unfortunately, we (usually) have to settle with a sample from the population.

Ideally, the sample is representative, allowing us to use probability and statistical inference to make conclusions that are generalizable to the broader population of interest.

Sampling methods

Probability sampling (e.g., simple random sampling, stratified, cluster, or multi-stage sampling)

- All units have a known chance of being selected
- More likely to be generalizable
- Can be more expensive and time-consuming

Non-probability sampling (e.g., quota, convenience, or snowball sampling)

- Some units unable to be selected, with no way of knowing size or effect of sampling errors
- Less generalizable to population of interest
- More convenient and less costly

Study design

Experimental studies (e.g., RCTs)

- Researchers directly control exposures or treatments
- Ability to make causal statements
- Less real-world applicability and generalizability

Observational studies (e.g., surveys)

- Researchers do not assign exposures or treatments
- Real-world setting with lower burden on participants
- Inability to prove causality

What can go wrong?

Selection bias, reporting bias, non-response bias, attrition bias, spin bias, confounding, detection bias, lack of blinding, straight up falsified data (this happens), ...

Catalogue of Bias CEBM UNIVERSITY OF OXFORD

...and so much more.

In recent news...

COVID-19 Antibody Seroprevalence in Santa Clara County, California

Eran Bendavid, Bianca Mulaney, Neeraj Sood, Soleil Shah, Emilia Ling, Rebecca Bromley-Dulfano, Cara Lai, Zoe Weissberg, Rodrigo Saavedra-Walker, James Tedrow, Dona Tversky, Andrew Bogan, Thomas Kupiec, Daniel Eichner, Ribhav Gupta, John Ioannidis, Jay Bhattacharya

doi: https://doi.org/10.1101/2020.04.14.20062463

This article is a preprint and has not been peer-reviewed [what does this mean?]. It reports new medical research that has yet to be evaluated and so should *not* be used to guide clinical practice.

(we'll be revisiting this study later this semester...)

In recent news...

Peer Review of "COVID-19 Antibody Seroprevalence in Santa Clara County, California"

Balaji S. Srinivasan Follow

Apr 17 · 14 min read ★

The high reported positive rate in this serosurvey may be explained by the false positive rate of the test and/or by sample recruitment issues.

What do some other people have to say about this study...?

Reproducibility and replicability

Reproducibility: being able to take the original data and code to reproduce all numerical findings

Replicability: being able to independently repeat an entire study without use of the original data (generally with the same methods)

Some best practices from the American Statistical Association:

- End-to-end scripting of research
- Use of version control and documentation
- Publication of code along with data

The current replication crisis

Drip, drip: Former Harvard stem cell researcher up to 18 retractions

Piero Anversa, a former star researcher at Harvard Medical School who left the institution under a cloud, is up to 18 retractions. But that's barely half of the 31 papers by Anversa's group that Harvard has requested journals pull over concerns about the integrity of the findings.

The two articles, published in the *Proceedings of the National Academy of Sciences*, appeared in 2008 and 2009. Anversa author, Annarosa Leri, are among the authors

Withdrawal: Maturation of lipop in the endoplasmic reticulum: C formation of functional dimers a aggregates.

Osnat Ben-Zeev, Hui Z. Mao and Mark H. Doolittle

VOLUME 277 (2002) PAGES 10727–10738

This article has been withdrawn by Osnat Ben-Zeev and Mark could not be reached. Fig. 3*A* contained several duplicated reg 6, *B* and *C*; 7, *A* and *B*; and 8, *A* and *B*, were inappropriately mar

© 2019 by The American Society for Biochemistry and Mol

RETRACTION | AUGUST 15, 2019

Kohrt HE, Houot R, Goldstein MJ, Weiskopf K, Alizadeh AA, Brody J, Müller A, Pachynski R, Czerwinski D, Coutre S, Chao MP, Chen L, Tedder TF, Levy R. CD137 stimulation enhances the antilymphoma activity of anti-CD20 antibodies. *Blood*. 2011;117(8):2423-2432.

Blood (2019) 134 (7): 658.

https://doi.org/10.1182/blood.2019002416 Article history

Connected Content

This is a retraction to: CD137 stimulation enhances the antilymphoma activity of anti-CD20 antibodies

The Editors of *Blood* retract the 24 February 2011 paper cited above. Concerns regarding the data underlying Figures 3A, 3B, 3C, 4A, and 5C were brought to the attention of Stanford University. The university investigated the issue, conducting a search for any original sources of these data. The search was unsuccessful; the experiments, data, and figure preparation for these figures were overseen by Holbrook E. Kohrt, who died before the university became aware of the concerns. As a result, the data underlying these figures cannot be validated.

THE UNITED STATES DEPARTMENT of JUSTICE

Home » Office of Public Affairs » News

JUSTICE NEWS

Department of Justice

Office of Public Affairs

FOR IMMEDIATE RELEASE Monday, March 25, 2019

Duke University Agrees to Pay U.S. $112.5 Million to Settle False Claims Act Allegations Related to Scientific Research Misconduct

Course syllabus

The course syllabus is the official document regarding all policies and guidelines and serves as the course syllabus. It is available on the course website here; a .pdf version is available on Sakai.