Statistical Science and Data Science

Nancy Reid

University of Toronto

11 July 2018

Statistical Inference

- data $\rightarrow$ conclusions
- data $\rightarrow$ uncertainty about conclusions

?data?

SHARE

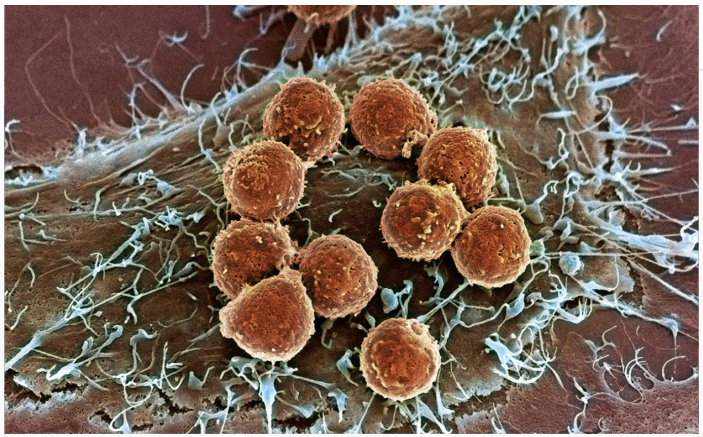
IN DEPTH | BIOMEDICINE

Blood test may predict cancer immunotherapy benefit

Ken Garber

+ See all authors and affiliations

Science 29 Jun 2018:
Vol. 360, Issue 6396, pp. 1387
DOI: 10.1126/science.360.6396.1387



"Detected (sensitivity [95% CI]) cancers (stage I-III) included 28 colorectal (66% [48-84]...) "

Garber, Science, June 28 2018
Klein et al., 2018 ASCO Ann. Mtg

Theory of Statistical Inference

- how to get from data to conclusions
- with generalizable strategies
- what principles do we use to develop these strategies
- how are these strategies to be evaluated — efficiency, precision
- a long history of the subject; using probability to both develop statistical methods and to evaluate their performance

  Bayes, Laplace, Gauss; Student, Fisher, Neyman, Pearson, Jeffreys, ...

- leading to confidence intervals, p-values, estimates and standard errors, etc.

  "66% [48 – 84]"

  "$p < 0.01$"

Role of Probability

- probability to describe physical haphazard variability — frequentist
  - probabilities represent features of the “real” world in somewhat idealized form
  - subject to empirical test and improvement
- probability to describe the uncertainty of knowledge — Bayesian
  - measures rational or “impersonal” degree of belief, — Jeffreys, 1939,1961

    or
  - measures a particular person’s degree of belief — F.P. Ramsey, 1926
  - linked to personal decision making

… role of probability

- confidence intervals or *p*-values refer to empirical probabilities

"66% [48 – 84]"

- inference is assessed by behaviour of the procedure under hypothetical repetition

- the Bayesian approach to inference describes uncertainty of knowledge
- this can be interpreted empirically by appeal to a notion of calibration

" … fundamentally false and devoid of foundation"

Big Data and Data Science

WIRED

2008

CHRIS ANDERSON SCIENCE 06.23.08 12:00 PM

THE END OF THEORY: THE DATA DELUGE MAKES THE SCIENTIFIC METHOD OBSOLETE

“faced with massive data, this approach to science — hypothesize, model, test — is becoming obsolete ”

“Petabytes allow us to say: ‘Correlation is enough’. We can stop looking for models”

Statistics and data science

JIM BRANDENBURG/MINDEN PICTURES/FLPA

Figure 1 | A glacier at Mount Robson Provincial Park, British Columbia, Canada. An analysis by Schildgen and colleagues[3] confirms that the rate of mountain erosion by glaciers has increased during the past few million years in certain places (such as in British Columbia) in response to climate cooling, but casts doubt on the idea that this was a global effect.

EARTH SCIENCE

Global erosion by glaciers revisited

Mountain erosion is thought to have sped up globally over the past few million years as the climate cooled and glaciers grew. A reassessment of the data suggests that this acceleration was limited to just a few regions. SEE LETTER P. 89

"Mountain erosion is thought to have sped up globally ... A reassessment of the data suggests that this acceleration was limited..."

Kirby, Nature July 2018

… Statistics and data science

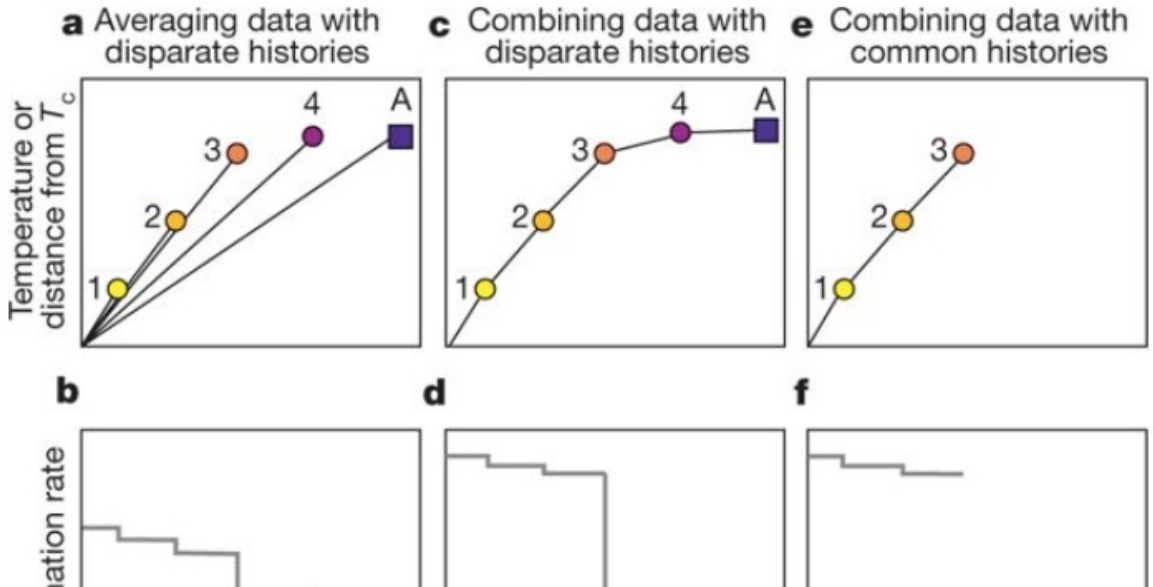
LETTER

https://doi.org/10.1038/s41586-018-0260-6

Spatial correlation bias in late–Cenozoic erosion histories derived from thermochronology

Taylor F. Schildgen[1,2,6]*, Pieter A. van der Beek[3,6], Hugh D. Sinclair[4] & Rasmus C. Thiede[2,5]

Fig. 2: Constructing erosion histories from multiple samples or multiple thermochronometers.



Statistics and data science

JIM BRANDENBURG/MINDEN PICTURES/FLPA

Figure 1 | A glacier at Mount Robson Provincial Park, British Columbia, Canada. An analysis by Schildgen and colleagues[3] confirms that the rate of mountain erosion by glaciers has increased during the past few million years in certain places (such as in British Columbia) in response to climate cooling, but casts doubt on the idea that this was a global effect.

EARTH SCIENCE

Global erosion by glaciers revisited

Mountain erosion is thought to have sped up globally over the past few million years as the climate cooled and glaciers grew. A reassessment of the data suggests that this acceleration was limited to just a few regions. SEE LETTER P. 89

"... as we use increasingly sophisticated analyses of **'big data'** to gain insight into global trends in geology, we must not lose sight of the physical processes that operate locally"

Kirby, Nature July 2018

… Statistics and data science

- combining data at multiple scales, or of multiple types
- complex dependencies — networks, images, text corpora, …
- looking for extremes — big data becomes small data

NEWS · 28 JUNE 2018

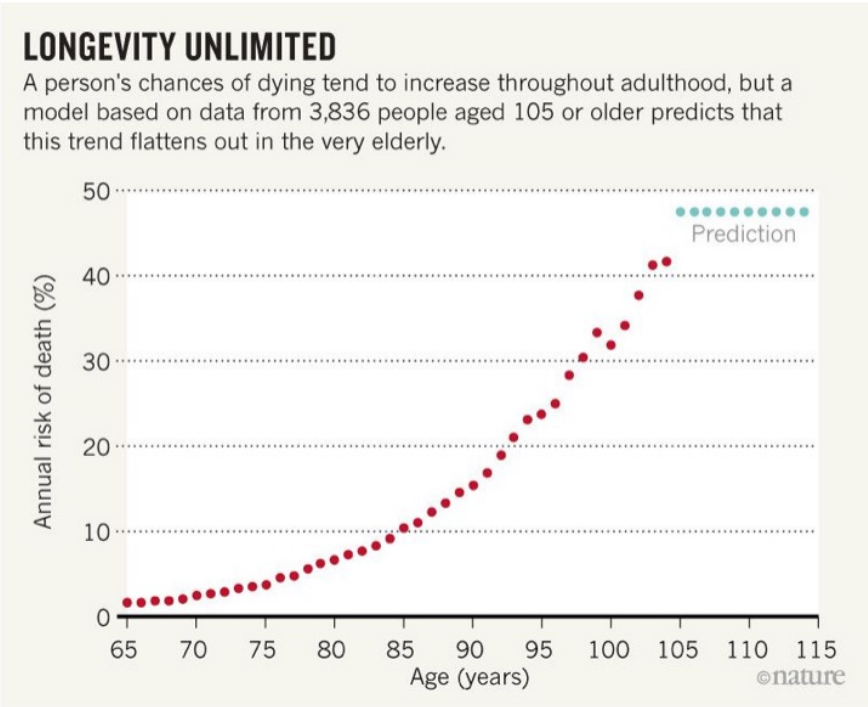
There's no limit to longevity, says study that revives human lifespan debate

Death rates in later life flatten out and suggest there may be no fixed limit on human longevity, countering some previous work.

Elie Dolgin

… Statistics and data science

- combining data at multiple scales, or of multiple types
- complex dependencies — networks, images, text corpora, …
- looking for extremes — big data becomes small data



"the study included fewer than 100 people who lived to 110 or beyond"

"even small inaccuracies in the Italian longevity records could lead to a spurious conclusion"

Dolgin, Nature, 28 June 2018

What is data science?

- short-hand for “lots of data”, “complicated data”, “data of uncertain provenance”
- an undergraduate or post-graduate program of training
- a job description
- a new multi-disciplinary field of study

- combining mathematics, statistics, computer science, domain science
- increased emphasis on privacy, fairness, communication, visualization, impact on policy, workflow and reproducible research

Blake & Olhede, 2016

Reproducibility

Is Most Published Research Really False?

Jeffrey T. Leek[1,2] and Leah R. Jager[1]

[1]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland 21205; email: jtleek@gmail.com

[2]Center for Computational Biology, Johns Hopkins University, Baltimore, Maryland 21205

ASA News
AMERICAN STATISTICAL ASSOCIATION
Promoting the Practice and Profession of Statistics®
732 North Washington Street, Alexandria, VA 22314 • (703) 684-1221 • Toll Free: (888) 231-3473 • www.amstat.org • www.twitter.com/AmstatNews

AMERICAN STATISTICAL ASSOCIATION RELEASES STATEMENT ON STATISTICAL SIGNIFICANCE AND *P*-VALUES

Provides Principles to Improve the Conduct and Interpretation of Quantitative Science

March 7, 2016

Statistical challenges in data science

- inference in very high dimensions ($p \to \infty$)

  requires dimension reduction or sparsity assumptions

- standard statistical methods may not scale well ($n \to \infty$)

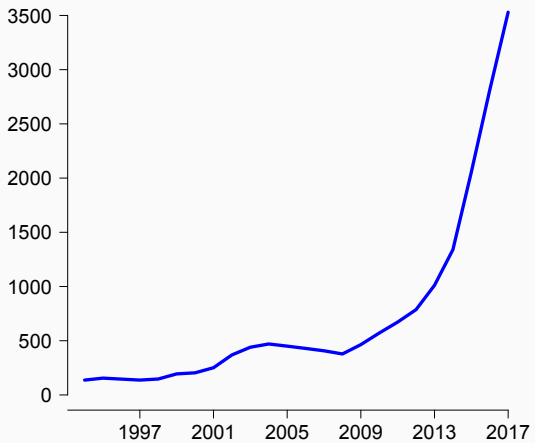
  distributed computing

- data provenance and quality

  "big data has arrived, but big insights have not" Harford, 2014

- learning causal relations from observational data *pace* WIRED

- data is everywhere requires rapid increase in training



2011: 672

2017: 3530

Those pesky *p*-values

- science is a process
- learning is incremental
- probability expresses uncertainty
- either epistemically or empirically

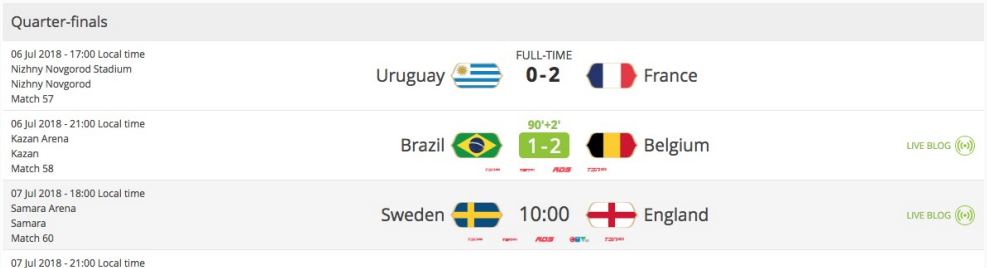
Figure 1 | A glacier at Mount Robson Provincial Park, British Columbia, Canada. An analysis by Schildgen and colleagues[1] confirms that the rate of mountain erosion by glaciers has increased during the past few million years in certain places (such as in British Columbia) in response to climate cooling, but casts doubt on the idea that this was a global effect.

EARTH SCIENCE

Global erosion by glaciers revisited

Mountain erosion is thought to have sped up globally over the past few million years as the climate cooled and glaciers grew. A reassessment of the data suggests that this acceleration was limited to just a few regions. SEE LETTER P.89

- for scientific advances, empirical behaviour of procedures is key
- for decision-making, personal probabilities have an important role



Quarter-finals

06 Jul 2018 - 17:00 Local time Nizhny Novgorod Stadium Nizhny Novgorod Match 57	Uruguay	FULL-TIME 0-2	France	
06 Jul 2018 - 21:00 Local time Kazan Arena Kazan Match 58	Brazil	90'+2' 1-2	Belgium	LIVE BLOG
07 Jul 2018 - 18:00 Local time Samara Arena Samara Match 60	Sweden	10:00	England	LIVE BLOG
07 Jul 2018 - 21:00 Local time				

Thank You!

Don Fraser, University of Toronto

David Cox, University of Oxford

Thank You!