

OPEN SOURCE

1

THE TRUTH MATTERS: ON THE RELATIONSHIP BETWEEN EVIDENCE AND CONCLUSIONS.

TAKING ANONYMOUS STATISTICAL LIVES AS SERIOUSLY AS IDENTIFIED INDIVIDUAL LIVES

There is a common preference to rescue and extend *named individual lives*, no matter what the cost.

Yet comparable investments might save millions of *anonymous invisible statistical lives*, since the cost extending a statistical life is often small compared to extending a named life.

named and nameless lives visible and invisible lives private vs. public interests
insiders and outsiders short vs. long time-horizons rescue treatments vs. prevention
personalized precision medicine (n = 1) vs. vaccination (n = 3,000,000,000)
local optimizing vs. global pessimizing proprietary vs. open source hedgehogs and foxes

*A mistake in the operating room can threaten the life of one patient,
a mistake in statistical analysis or interpretation can lead to hundreds of early deaths.
So it is odd that, while we allow a doctor to conduct surgery only after years of training,
we give software packages – SPSS [Python, R, MATLAB, Machine Learning, et al] –
for statistical analysis to almost anyone.* ANDREW VICKERS

THE STANDARDS OF STATISTICAL REASONING ABOUT THE TRUTH ARE UNIVERSAL

*Increasing knowledge begets increasing specialization and narrower scope of understanding.
Statistics, as the practice of planning experiments and observations and of interpreting data, has a common
relation to all sciences. Unification will be more easily attained if the logical framework of the individual
sciences can be identified and isolated from their factual content.* FRANCIS ANSCOMBE, ET AL.

*A statistician I knew once replied to a brain surgeon whose hobby was statistics
that hers was brain surgery.* STEPHEN JOHN SENN

*Although we often hear that data speak for themselves, their voices can be soft and sly.
It is easy to lie with statistics; it is easier to lie without them.* FREDERICK MOSTELLER

*Confirmation bias is the tendency to search for, interpret, favor, and recall information so as to confirm
one's pre-existing beliefs or hypotheses.* WIKIPEDIA *It is a principle that shines impartially on the just and
the unjust alike that once you have a point of view all history will back you up.* VAN WYCK BROOKS

*Bias can creep into the scientific enterprise in all sorts of ways. But financial conflicts are detectable
definitively and represent a uniquely perverse influence on the search for scientific truth.* COLIN BEGG

*It is simply no longer possible to believe much of the clinical research that is published, or to rely on
the judgment of trusted physicians or authoritative medical guidelines. I reached this conclusion
slowly and reluctantly during my 2 decades as an Editor of The New England Journal of Medicine as
drug companies asserted more power, and began to treat researchers as hired hands.* MARCIA ANGELL

WHEN THE TRUTH MATTERS: THE CREDIBILITY OF INFERENCES

How do ^{you} I know that?
they

How could ^{you} I possibly know that?
they

To reason about evidence and conclusions, to evoke self-awareness about the truth of our knowledge, to measure explanatory depth, ask *How do I know that? How do we know what we don't know?* Feynman's law: 'The first principle is that you must not fool yourself, and you are the easiest person to fool.' Ask others *How do you know that?* Then, thought experiments about knowledge: *How could anyone possibly know that? What research design would produce credible evidence for the claimed knowledge?* If none, the claimed knowledge is not even wrong – for it is impossible to prove or disprove.

The fundamental principles of analytical thinking: reason about causality and mechanism, explain something, make comparisons, assess credibility of measurements and inferences, validate findings, enforce integrity and honesty. These *principles cannot be altered or repealed* by assumptions, by any discipline or specialty, political and intellectual fashions, marketing or monetizing, what sponsors desire or demand, by anything you do, think, believe, hope.

WHY RESEARCH ON HUMANS IS WAY MORE DIFFICULT THAN ROCKET SCIENCE

Nature's mathematical laws apply to every particle everywhere forever. From tiny particles and undetectable waves up to the entire universe, all measurements testify about those laws. Truth and exactitude are always present – and so real science is easy, compared to understanding biological systems and human behavior, which lack mathematical laws.

In biological systems, DNA masterplans are a cumulative tangle of local random evolutionary hacks, work-arounds, mutations. Unlike universal physical laws, 'living organisms are historical structures, literally creations of history. They represent a patchwork of odd sets pieced together when and where opportunities rose.' Some biological complexities appear unfathomable and irredeemable: 'random replicative mutations in stem cell divisions (bad luck) are largely responsible for variations in cancer risk.'

Research on humans involves the space-time mash-ups of biological systems – and humans who act, plan, think, connive, cheat, think about thinking, and fail to recognize their ignorance. Leo Tolstoy's *War and Peace* described the difficulties of making causal inferences about human behavior:

'When we say that Napoleon *commanded* armies to go to war, we combine in one simultaneous expression a whole series of consecutive commands dependent on one another. Our false idea that an event is caused by commands that precede it, out of 1000s of commands those few that were executed and consistent with the event – and we forget about the others that were not executed because they could not be.'

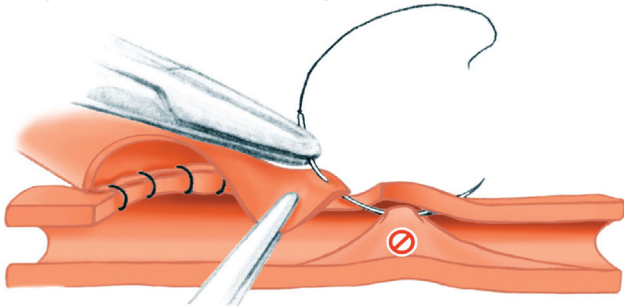
With greater understanding of human/biological complexities, life expectancy has *doubled globally since 1900*. In 1950, worldwide life expectancy was 48 years, and in 2019, 71 years. Advances in public health, science, economic development, education, and evidence-based medicine produced these unprecedented gains in human history. Then came the pandemic.

**REMODELING DATA MEASUREMENT AND ANALYSIS:
TAKING STATISTICAL LIVES AS SERIOUSLY AS INDIVIDUAL LIVES.**

Cardiac Surgery: Safeguards and Pitfalls in Operative Technique compiled by Siavosh Khonsari and Colleen Sintek is an intense encyclopedia of 3,000 alerts and warnings:

🚫 = avoid this grave error, how to prevent it, and if it occurs, what to do,
and the less severe alerts ⚠️ = warning, **NB** = note well. Here is one alert from 3,000:

A needle inadvertently picking up the bottom of an artery wall



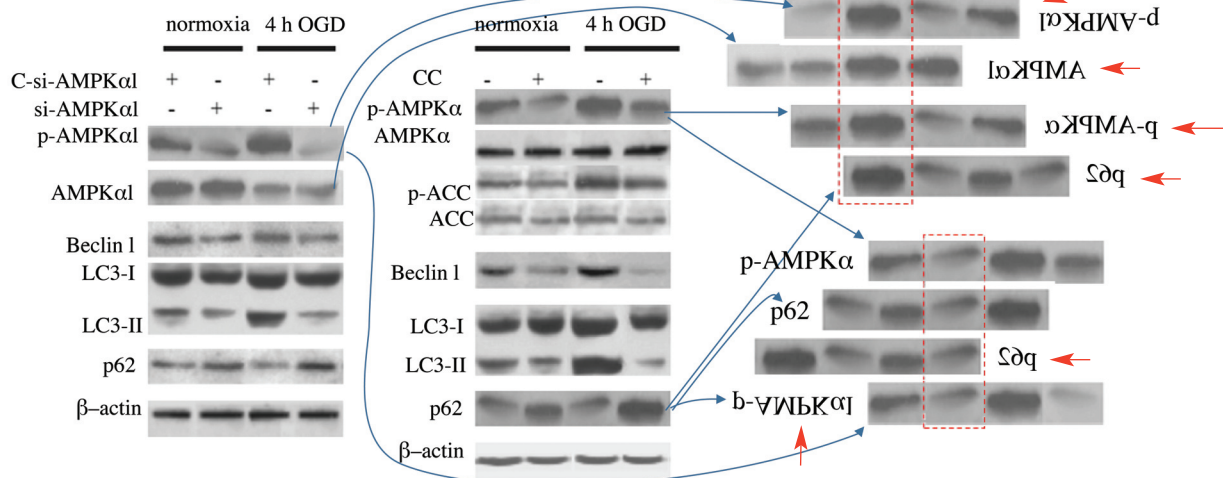
🚫 INADVERTENT SUTURE OF POSTERIOR WALL

“The anastomosis toe is the most critical part because it determines graft outflow capacity. When the artery lumen is small or visibility is suboptimal, the needle may pick up the artery posterior wall. An appropriately sized plastic probe passed for a short distance into the distal artery allows precise placement of sutures and prevents occurrence of this complication.”

Imagine now another encyclopedia, *Data Measurement and Analysis: 1,000 Safeguards, Pitfalls, and Cheats in Statistical Practice*, where quality control for statistical lives would seek to match quality control for named lives in cardiac surgery. An encyclopedia entry:

🚫 INAPPROPRIATE IMAGE DUPLICATION

In research papers, fraud is often detected by carefully examining images and data graphics. In this retracted paper, apparently falsified Western Blot Tests were created by copying nearby blots, and then disguised by flipping and mirror-reversing. This problem is obvious because the blot labels were inadvertently flipped and reversed, resembling upside-down Russian words!



PREVALENCE OF MEASUREMENT AND DATA ANALYSIS PITFALLS IN MEDICAL RESEARCH

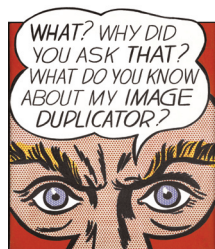
Examining 1000s of published studies, meta-researchers measure prevalences of pitfalls and cheats in data collection/analysis. The prevalence rates are appalling. (Should we be thankful that 95% of published medical research vanishes, unread and uncited even by the authors' mothers?) For example, early reports of medical interventions make enthusiastic claims that never again will be achieved. As evidence improves, harms may eventually exceed benefits (eg, baby aspirin for heart attack prevention) – and prevalence of regression toward the truth is perhaps 80%. Here are prevalence rates of major problems in medical research: randomization failures, inability to undo compromised research in a timely way, image integrity, measurement quality – more entries for a proposed encyclopedia *Data Measurement and Analysis: 1,000 Safeguards, Pitfalls, and Cheats in Statistical Practice*.

🚫 COMPROMISED RANDOMIZING: THE MEDITERRANEAN DIET STUDY AND ITS 267 OFFSPRING

🚫🧠🌟 RETRACTION SYNDROME 🌟 INDEPENDENT AUDITS OF RESEARCH = A NECESSITY

'Precise fulfillment of randomization protocols assigning subjects to treatment vs. control is critical to research design. Imprecise randomization causes serious complications,' reports the thorough John Carlisle, who detected incorrect randomizing in the famous Mediterranean Diet study (NEJM 2013). This was retracted, corrected data analyzed and published – with somewhat weaker evidence for Mediterranean diets (NEJM 2018). 267 secondary articles based on original incorrect data still remain at large. Prevalence of randomizing mess-ups are based on 5,087 articles; note strongly-worded titles concerned about credibility and truth: "Data fabrication and other reasons for non-random sampling in 5,087 randomized, controlled trials in anesthetic and general medical journals," John B. Carlisle, *Anesthesia* 72 (2017), 931-935. "PREDIMED trial of Mediterranean diet: retracted, republished, still trusted?" Arnav Agarwal and John P.A. Ioannidis, *BMJ*, 7 Feb 2019.

🚫 INAPPROPRIATE IMAGE DUPLICATION: 3.8% AND 6.1% PREVALENCE RATES



"Images from 20,621 papers published in 40 scientific journals 1995-2014 were visually screened: 3.8% of published papers contained problematic images, half exhibited features suggesting deliberate manipulation." Another study reviewed 960 papers from *Molecular and Cellular Biology* 2009-2016 and found 6.1% (59 of 960) "papers to contain inappropriately duplicated images, leading to 41 corrections, 5 retractions."

"Prevalence of inappropriate image duplication in biomedical research publications," Elisabeth Bik, Arturo Casadevall, Ferric Fang, *ASM mBio* 7, 2016; "Analysis and correction of inappropriate image duplication," Elisabeth Bik, et al *MCB* (2018).

🚫 GENE NAME CONVERSION ERRORS IN EXCEL: 704 ARTICLES IN 18 GENOMICS JOURNALS

In default settings, Microsoft Excel converted gene names with 3 letters and 2 numbers to dates and floating-point numbers. "A programmatic scan of leading genomics journals reveals 20% of papers with supplementary Excel gene lists contain erroneous gene name conversions." This error was known to insiders in ~2004; alas from 2005-2015, 704 papers published in 18 journals were affected. "Gene name errors widespread in scientific literature," Mark Ziemann, Yotam Eren, Assam El-Osta, *Genome Biology* 17 (2016). Update: James Vincent, "Scientists rename human genes to stop Microsoft Excel from misreading them as data: Sometimes it's easier to rewrite genetics than update Excel," *The Verge*, August 6, 2020.

FACING UP TO MEASUREMENTS

☆☆☆ OBSERVE DATA COLLECTION AT THE MOMENT OF MEASUREMENT

See, observe, learn how data are collected at moment of measurement. “You never learn more about a process than when you directly observe how data are actually measured,” said Cuthbert Daniel, a superb applied statistician. See with fresh eyes. Walk around what you want to learn about. Talk to those who do measurements. See how numbers came to be. Are those measuring skilled/honest/biased/incompetent/tired and emotional/sloppy? Are errors and artifacts in measurements assessed? How are outliers adjudicated? Those directly observing medical measurements may well rightly conclude that medical care/research is a two-digit science on a very good day. On other days, getting the sign right is an achievement.

Example: big company polluted big river, environmental agencies forced polluters to clean up and monitor progress by daily water samples. Small boat goes out to take sample, dipping into the water *after looking around for the cleanest water*. Statisticians call this ‘sampling to please.’ Observing this data collection reveals the early limits of self-monitoring, and that people can’t keep their own score.

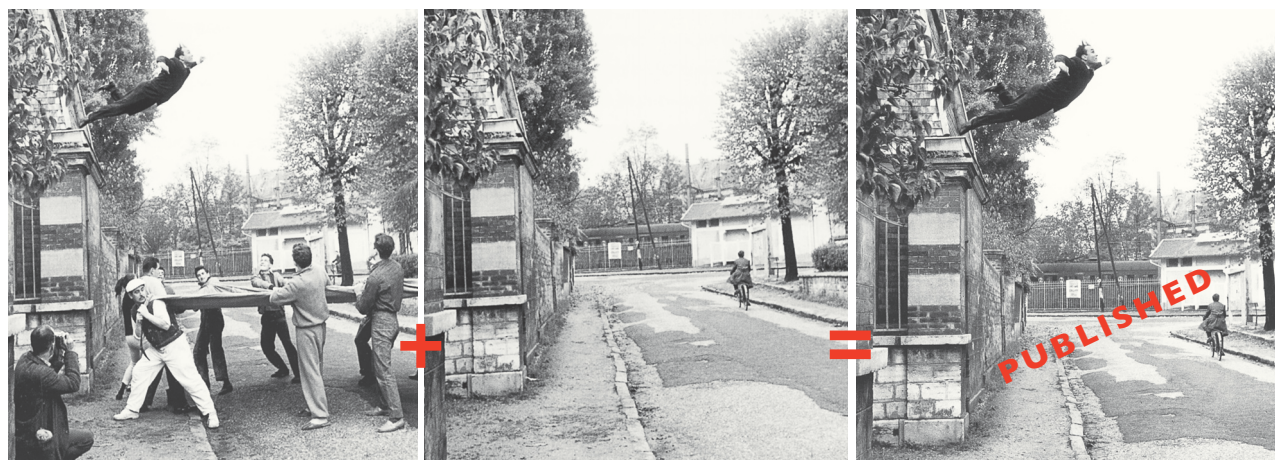
Read service manuals on measurement practices and their artifacts. Do detective work. Nurses and techs make measurements all day long; ask about quality, errors, relevance, false alarms, and duplicate/unnecessary/lucrative measurements that signal over-diagnosis. People love to talk about their work, let them do the talking. Ask others the same questions. Check, verify, recheck reports. From 34 hours (2019-2021) of my interviews with ICU/oncology nurse (U.S.):

Q. End-of-life care in practice? A. “Oncologist pulled near-death patient out of hospice who reportedly ‘looked better.’ What did patient do, smile – vitals are not even measured in hospice! Patient sent to oncology floor, I refused to administer chemo because pt was about to die, oncologist backed down when I threatened to inform Patient Ethics Committee, pt died that night.” Q. With your additional responsibilities as floor nurse-in-charge, how much more you paid? “\$1.00 per hour at both hospitals where I’ve worked.” Q. How can that be?! “94% of nurses are females” Q. Doctor quality? “Big range, a few surgeons are awful, everyone knows who, some anesthesiologists won’t work with them.” Q. Cause of Death forms? [long discussion = death adjudication is uncertain]

Q. (On covid, Spring, 2021 interviews): “Nurses the only people seeing patients. Proning, placing pt stomach down, reduced ventilator use. Hospital supplied one N95 mask for 2 weeks reuse. PPE shortage. No hazard pay, they said medical center ‘had lost \$100s millions.’ 17 nurses on my floor got covid-19.” Q. Biggest thing you learned? “[Hospital Company] does not care about nurses.” Q. What’s new? “For Nurses week, Medical Center gave Nurses a rock in paper bag, note saying ‘make a Gratitude Rock, paint this, give to a friend, express your gratitude.’”

☆☆☆ OBSERVE DATA AT MOMENT OF MEASUREMENT

NOT AT MOMENT OF PUBLICATION



☆☆☆ TRACK THE ORIGINS AND LIFE HISTORIES OF MEASUREMENTS

Why is the measurement made? Who looks at the measurements, when, where, why?
 What are consequences, harms, benefits? Where did the data go, what's it doing now?
 For medical measurements, track the money: who profits how much? Behind the scenes,
 engineering/technical manuals frankly describe measurement complexities and problems.

☆☆☆ ACTUAL MEASUREMENTS VS. PUBLISHED DATA: INVESTIGATION OF SCIENTIFIC MISCONDUCT

FROM "THE REPORT OF THE INVESTIGATION COMMITTEE ON THE POSSIBILITY OF SCIENTIFIC MISCONDUCT IN THE WORK OF HENDRIK SCHÖN AND COAUTHORS, APPENDIX E: ELABORATED FINAL LIST OF ALLEGATIONS." LUCENT TECHNOLOGIES, SEPTEMBER 2002. EDITED.

"Can the data presented be traced back to primary data, free of any data processing or other manipulation?"

It is a well-established tenet of science that clear records should be kept. Only credible, primary data can provide unambiguous corroborating evidence for published data. An understanding of the procedures of data acquisition and analysis also provides a context within which possibly mitigating circumstances can be assessed. It is worth emphasizing that the retention of primary data, together with adequate record keeping, are necessary to the ordinary conduct of science, not simply for the examination of possible wrongdoing. In the conduct of research, new questions arise that require a revision of the original analysis, and thus require a return to the primary data. Failure to keep primary data and records for a reasonable time is, by itself, a threat to the health of the scientific enterprise. This remains as true in the computer age as it has been in the past.

Is there clear evidence that the data do not come from the measurements described?

This evidence takes different forms: *Data Substitution*, in which data sets for distinct experimental conditions show unreasonable similarity to each other, in some cases after multiplying one data set by a constant factor; *Unreasonable Precision*, in which a data set agrees better with a simple analytic expression than would be expected from the measurement accuracy; and *Contradictory Physics*, in which the data appear to be inconsistent with prevailing scientific understanding and description of the measurement. Many great discoveries in science would at first have been included in the Contradictory Physics category, so the Committee has set aside all but a few especially problematic examples. But extraordinary results demand extraordinary proof. Unless special diligence is demonstrated, results that contradict known physics suggest simple error, self-deception, or misrepresentation of data.

If the data are not valid, are there mitigating circumstances that explain how the data came to be misrepresented?

For example, a clerical error in including the wrong data in a figure represents poor procedures, but not misconduct. But such innocent explanations may require understanding of the state of mind of the authors at the time the data were prepared, and this cannot be determined definitively. It must be noted that the credibility of a particular innocent explanation depends on the overall credibility of the scientist in question. This in turn depends on whether there is an unreasonable number of problems or a pattern of questionable practices. Rather, the problems with the data are already established, and the question is whether many improbable, innocent explanations should be accepted."

☆☆☆ 'YOU CAN OBSERVE A LOT BY JUST WATCHING. OF COURSE YOU DO HAVE TO KNOW WHAT TO WATCH, AND YOU DO HAVE TO KEEP WATCHING'

Cuthbert Daniel: "The name 'statistician' meant something different to management from what it meant to me. For them, it was complete records of what the plant was doing at all points, all times: a graph of U₂₃₅ concentration versus Building Numbers from 1 to 47 was a curve that went from 0.7% (where it starts) up to >90%. One of my jobs was to keep weekly records of average measurements made by mass spectroscopy in each building. After a month I noted while the graph went steadily up, one building showed flatness, not a step. Instead of rushing to management saying something is wrong in building 41, I went to the smartest Process Engineer and said, 'If a building wasn't working at all, how would it show?' He said, "That's easy. Valve F43 is open, bypassing the barriers.' Then I was alarmed, went to the plant manager, said 'Valve F43 has been open for the last month.' That's the only time I ever got any attention from him and I got it then. How did I know that? *It was a graph, no exact numbers, no math.* Only weekly building averages. This exemplifies the Yogi Berra Principle: 'You can observe a lot by just watching.' Of course you do have to know what to watch, and you do have to keep watching!" ET, "Conversation with Cuthbert Daniel,"

Statistical Science, 1988, 413-424

⚠ ARE MEASUREMENTS REAL, OR JUST BANG-BANG DUPLICATES?

An observer measures and records a number and then a few seconds later records another. These two measurements are not independent if the process changes hourly. Knowledge of the first number gives full knowledge of the second, and the number of independent measurements is not two but one. In econometrics, this is called *autocorrelation* or *serial correlation*; in *experimental design*, *bang-bang duplicates* or *pseudo-replications*. Duplicates do show up in mouse studies: if a researcher does an experiment on 3 mice and measures the same variable on each mouse 30 times, the sample size is not 90. The only way to achieve statistical significance with 3 mice is if one of them turns into a cat.



⚠ CONSEQUENCES OF A SPREADSHEET FILLED WITH NOISE



Spreadsheets of random numbers contain no viable findings, except that you've got the wrong spreadsheet. Dr. Confirmation Bias and Dr. pHacker fabricate stories based on noise. WEIGHTED RANDOM NUMBER means that while some sports competitors are better than others, the exciting fine-grain local variations around averages are random, luck, coincidences, cheats, outliers, miracles. Weak evidence spawns big attitudes, a rage to conclude: "Ignorance more frequently begets confidence than does knowledge," said Charles Darwin.

⚠️ YOUR DATABASE MAY NOT CONTAIN THE TRUTH OR THE ANSWERS TO ALL OF YOUR QUESTIONS

Is the database at hand capable of answering the research questions at hand? A database may not contain the relevant explanatory variables, a devastating constraint. And why should any data set contain all relevant known knowns and unknown knowns, and known unknowns, and the most difficult, unknown unknowns? Many data analyses are model specification searches – let's try this, let's try that. But *this* and *that* might not even be in the database. Get more/different data. Far better, have an independent explanatory theory, as in real science. Truth = explanatory theory, evidence, independent replication, and quality/honesty/integrity in conducting research. Enemies of the truth: (1) no explanatory theory, (2) no empirical truth available for the topic at hand, (3) lack of relevant data, (4) incompetence, gullibility, lack of self-skepticism, (5) Drs. Confirmation Bias and their conflicts of interest, (6) cheats, lies.

⚠️ 97% OF 565 MEDICAL RESEARCH REPORTS FAILED TO DEAL WITH MEASUREMENT ERRORS

A 2018 analysis examined practices in reporting measurement errors in 12 major medical and epidemiology journals: 44% of research articles mentioned measurement errors, and only 7% of those investigated or corrected the errors.

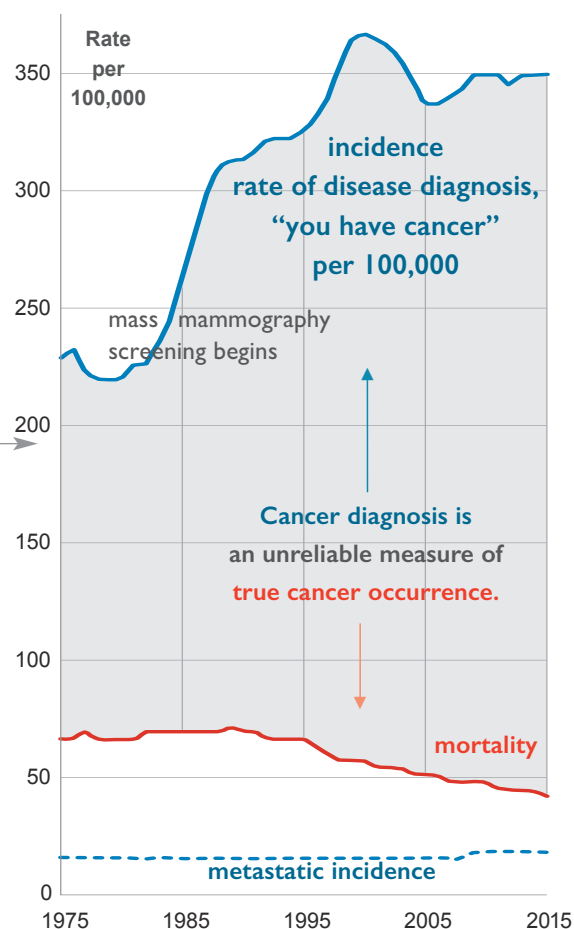
Timo B. Brakenhoff, Marian Mitroiu, Ruth H. Keogh, Karel G.M. Moons, Rolf Groenwold, Maarten van Smeden, "Measurement error is often neglected in medical literature," *Journal of Clinical Epidemiology*, 2018.

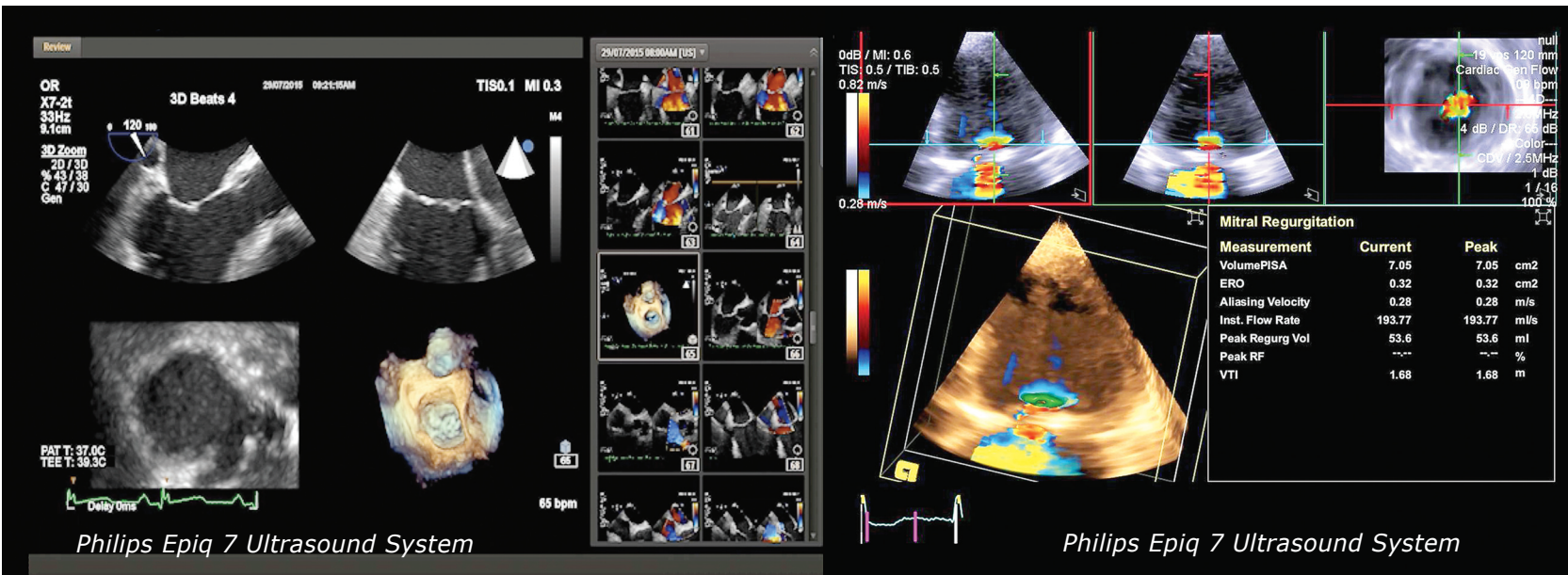
⚠️ IN HEALTHCARE, FALSE ALARMS ARE THE MOST PREVALENT AND LUCRATIVE MEASUREMENT ERRORS

Screening tests produce many false alarms, terrifying millions of healthy people. False alarms cascade into more tests. Mass screenings are now regarded as dubious – because of false alarms, harms, and failure to reduce all-cause mortality. Early diagnosis leads to early cures, or that patients just get the bad news sooner. Since survival time = time from diagnosis to death, early diagnosis can create statistical illusions of improved survival times. And false alarms, if their falsity is not detected, lead to treatments of patients for a disease they don't have. This entire gray area shows a mix of over-diagnosed, cured, indolent, incidentals, subclinical, and harmless cancers (many older people die *with* cancer, not *of* cancer). Data on the Number Needed to Treat (the number of patients needed to treat for a single favorable outcome) indicates that from 2 to 1000 cancer patients are treated for each one that benefits. Often far more are harmed than benefitted.

Data from 2014–2015 Surveillance Epidemiology End Results (SEER). Gilbert Welch, Barnett S. Kramer, William C. Black, "Epidemiologic Signatures in Cancer," *New England Journal of Medicine*, Oct 2019, redrawn. See also Andrea R. Marcadis, Jennifer L. Marti, Behfar Ehdai, et al, "Characterizing Relative and Disease-Specific Survival in Early-Stage Cancers," *JAMA Internal Medicine*, December 9, 2019.

Breast cancer in women ≥40 years of age





**REAL-TIME DATA ANALYSIS WHEN THE TRUTH MATTERS:
MULTIPLE 2D/3D MEASUREMENTS DESCRIBE PATIENT STATUS AND
PERFORM LIVE QUALITY CONTROL DURING HEART SURGERY**

On dividing a narrow space between the aorta and pulmonary artery:

Senior heart surgeon: "Divide exactly in the middle."

Student of heart surgery: "If I err, to which side should I err?"

Senior heart surgeon: "Don't err." DR. MARC GILLINOV

An excellent data visualization, transesophageal echocardiograph (TEE, pronounced "T-E-E") provides *live imaging and quantitative measurements during heart surgery, exactly at point of need*. Requiring substantial real-time signal-processing, TEE imaging is based on ultrasound data from a transducer/receiver placed within a few centimeters of the heart by going down the patient's esophagus, bouncing ultrasound waves off the heart, then converted to images and data.

When the heart surgery is almost completed, but the patient is still opened up, the surgical results up to this point are evaluated: the patient is taken off the heart-lung machine, the heart restarted, then TEE provides visual assessments on how the newly repaired heart is working. If an issue appears, the patient goes back on-pump, heart stopped again, issue repaired, patient then closed up. TEE images are displayed on a low-glare matte screen for surgeon and sonographer, and shown on two large glary display screens viewed by the 12 people in the OR. This is a superb visualization and data model: it stays close to the data, processes and analyzes data in real-time, leads to quality control and interventions, works hand-in-hand with the surgical team. (In some other medical situations, TEE is overused with few diagnostic benefits.)

EXCELLENT ASSESSMENT OF ARTIFACTS AND ERRORS IN 2D/3D TEE DATA

TEE technical manuals describe how 2D and 3D imaging models produce artifacts in daily use. Sonographers can identify image artifacts and then make corrections by adjusting the position of the ultrasound transducer/receiver. Note the attention to detail. *This is what serious empirical analysis of measurement errors looks like, and it is enormously better than the models of "error" in classical statistics.*

Look over next page with care, just read around the technical jargon.

“Acoustic Artifacts in 2D Imaging Acquisition

The transducer adds noise to the echo signal by beam-width effects, axial resolution limitations, frequency characteristics. Control choices made by sonographers affecting amplification, signal processing, and echo signal can lead to significant differences in echo data. **Acoustic saturation** occurs when received signals reach a system's high-amplitude limit, when the system is unable to distinguish or display signal intensities. At saturation, increased input will not increase output. **Aliasing** occurs when the detected Doppler frequency exceeds the Nyquist limit. On the spectral display by Doppler can show peaks going off the display top or bottom, wrapping around the other side of the baseline. On the color display an immediate change in color from one Nyquist limit to the other is seen. **Comet tail** is a reverberation artifact produced when two or more strong reflectors are close together and have a high propagation speed. Then, sound does not travel directly to a reflector and back to the transducer; a strong linear echo appears at the reflector and extends deeper than the reflector. **Enhancement** is an increased relative amplitude of echoes caused by an intervening structure of low attenuation. **Focal enhancement or focal banding** is increased intensity in the focal region that appears as a brightening of echoes shown on the display. **Mirror imaging artifact** is commonly seen around the diaphragm; this artifact results from sound reflecting off another reflector and back. **Mirroring** is the appearance of artifacts on a spectral display when there is improper separation of forward and reverse signal processing channels. Consequently, strong signals will mirror into the other. **Multi-path positioning** and **refraction** artifacts takes place when paths to and from a reflector are different. The longer the sound takes traveling to or from a reflector, the greater the axial error (increased range) in reflector positioning. Refraction and multi-path positioning errors are normally relatively small, contributing to general image degradation rather than to gross errors in object location. **Propagation speed errors** occur when the assumed value for propagation speed by the ultrasound system is incorrect. If the actual speed is greater than assumed, the calculated distance to a reflector is too small, and the reflector will be displayed too far from the transducer. Speed error can cause a structure to be displayed with incorrect size and shape. **Range ambiguity** occurs when reflections are received after the next pulse is transmitted. In ultrasound imaging, it is assumed that for each pulse produced, all reflections are received before the next pulse is sent out. The ultrasound system calculates distance to a reflector from the echo arrival time assuming all echoes were generated by the last emitted pulse. The maximum depth to be imaged unambiguously by the system determines its maximum pulse repetition frequency. **Reverberation** is the continuing reception of a particular signal because of reverberation rather than reflection from a particular acoustic interface. Reverberations are easily identifiable, because they are equally spaced on the display. **Scattering** is the diffuse, low-amplitude sound waves that occur when acoustic energy reflects off tissue interfaces smaller than a wavelength. In diagnostic ultrasound, Doppler signals come primarily from acoustic energy back-scattered from red blood cells. **Shadowing** is the reduction in echo amplitude from reflectors that lie behind a strongly reflecting or attenuating structure. This phenomenon occurs when scanning a lesion or structure with an attenuation rate higher than that of the surrounding tissue. The lesion causes a decrease in beam intensity, which results in decreased echo signals from the structures beyond the lesion. Consequently, a dark cloud behind the lesion image forms on the display. This cloud, or shadow, is useful as a diagnostic clue. **Side lobes** (from single-element transducers) and **grating lobes** (from array transducers) cause objects that are not directly in front of the transducer to be displayed incorrectly in lateral position. **Speckle** artifacts appear as tissue texture close to the transducer but does not correspond to scatterers in tissue. It is produced by ultrasound wave interference and results in general image degradation. **Spectral broadening** is a display phenomenon that occurs when the number of energy-bearing Fourier frequency components increases at any given point in time. As a consequence, the spectral display is broadened. Spectral broadening can indicate disturbed flow caused by a lesion, and it is important diagnostically. However, broadening can also result from interaction between flow and sample volume size. **Speed of sound** artifacts occur if the sound propagation path to a reflector is partially through bone, where sound speed is greater than in average soft tissue. Reflectors appear closer to the transducer than their actual distance because of this greater speed of sound, resulting in a shorter echo transit time than for paths not containing bone.

Acoustic Artifacts in 3D Imaging

Acquisition, rendering, and editing artifacts are specific to 3D volume images. Acquisition artifacts are related to patient motion, organ motion, or position-sensing errors. Rendering artifacts include elimination of structures by limiting the region of interest boundaries, thresholding that eliminates structures, and adjacent structure artifacts that add additional information or hide structures. Editing artifacts result from data deleted from a rendered image. **Color** and **Color Power Angio** artifacts include a color flash artifact occurs when gain is set high and the transducer or patient moves. When gain is set too high, the color ROI box fills with color flash. When gain is set low, color bleed can occur. When gain is set too low, insufficient color data renders the image undiagnosable. **Color gain, directional, and motion artifacts** occur in 3D imaging. Color gain artifacts are mainly related to the use of excessive gain resulting in random color patterns in 3D images that might be interpreted as diagnostically significant. Directional artifacts are due to aliasing or directional confusion. The velocity range must be set properly, and the relationship between the transducer orientation and the flow vector must be understood. Patient motion can produce flash artifacts that are less obvious in 3D images than in 2D. **Dropout and shadowing** are present in 3D imaging although they are more difficult to recognize due to different and unfamiliar displays. Acoustic shadowing and other artifacts look very different when displayed in 3D volumes and may be more difficult to recognize than on standard 2D imaging. Those artifacts may produce apparent defects, such as nonexistent limb abnormalities or facial clefts. Acquiring data from multiple orientations may these artifacts. Fetal limb deficit artifacts are specific to 3D volume images. One explanation for the missing limbs is shadowing caused by adjacent skeletal structures. Overcoming the limb deficit artifact can be accomplished by changing the transducer position and the acquisition plane, as usual. **Motion artifacts** in 3D volumes can be caused by patient motion, fetal movement, cardiac motion, and other movement. Patient motion can produce flash artifacts that are more obvious in 3D images than in 2D. **Pseudoclefting and artifacts** are similar to limb deficit artifacts. Artifacts may be present in 3D imaging of the fetal face. As with 2D imaging, it is important to verify putative physical defects by using additional images and other modalities. **Resolution, attenuation, and propagation artifacts** are common to 3D imaging. Careful scrutiny of original 2D images is necessary to identify/preclude these artifacts from 3D volume image.”

☆☆☆ THE EFFECTIVENESS OF APPROXIMATE MEASUREMENTS AND MODELS

Frederick Mosteller's brilliant essay on approximate vs. refined measurements – and their relation to policy interventions.

“It is the experience of statisticians that when fairly ‘crude’ measurements are refined, the change more often than not turns out to be small. Merely counting laboratories in a school system is a crude measurement. It is possible to learn more about the quality of laboratories [and test our skepticism about the original crude measurements].

But statisticians would not leap too readily to that . . . Sadly, in real life the similarities of basic categories are often far more powerful and important than the nice differences which can come to absorb individuals so disposed, but which really don't make a great difference in the aggregate. Statisticians would wholeheartedly say make better measurements, but they would often give a low probability to the prospect that finer measures would produce data leading to different policy. The reasons are several. One is that policy decisions are rather insensitive to the measures – the same policy is often good across a great variety of measures. Secondly, finer measures are something like weights. For example, perhaps one science laboratory is only half as good as another – well and good, let us count it $1/2$. It turns out as an empirical fact that in a variety of occasions, we get much the same policy decisions in spite of weights. So there are technical reasons for thinking that finer measurement may not change the main thrust of one's policy. None of this is an argument against getting better information if it is needed, or against having reservations. More data cost money, and one has to decide where good places are to put the next money acquired for investigations. If we think it matters a lot by all means let us measure it better.

Note distinctions between statistical lives and individual lives. For schools, we may not know what works well for specific students, but we do know that if we stopped teaching algebra, few people would ever learn algebra.

Another point about aggregative statistics is worth emphasizing for large social studies. Although the data may not be adequate for decisions about individual persons, they may well be adequate for deciding policy for groups. We may not be able to predict which ways of teaching spelling will be preferable for a given child, but we may be able to say that, on the average, a particular method does better. And then the policy is clear, at least until someone learns how to tell which children would do better under the differing methods.”

APPROXIMATE MODELS

Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise. JOHN TUKEY

Simple methods typically yield performance almost as good as more sophisticated methods, to the extent that the difference in performance may be swamped by other sources of uncertainty that generally are not considered in the classical supervised classification paradigm. DAVID HAND

★ **PRIOR TO DATA ANALYSIS, CONDUCT AN INDEPENDENT FORENSIC DATA AUDIT OF YOUR DATA. DON'T JUST LOOK AROUND, INSTEAD SEE EVERYTHING**

If the truth matters, spreadsheets require unbiased forensic audits. Audits are sometimes deflected because researchers and sponsors are anxious to get a peek at the findings early on, despite the well-known anchoring bias that early information has undue influence. Chris Groskopf's *Guide to Bad Data* reveals 46 data quality issues in spreadsheets, a good start for forensic audits.

“Values are missing Zeros replace missing values Data are missing you know should be there
 Rows or values duplicated Spelling inconsistent Name order inconsistent Date formats inconsistent
 Units are not specified Categories badly chosen Field names ambiguous Provenance not documented
 Suspicious values present Data too coarse Totals differ from published aggregates
 Spreadsheet has 65,536 rows Spreadsheet has 255 or 256 columns Spreadsheet has dates 1900, 1904, 1969, 1970
 Text has been converted to numbers Numbers have been stored as text
 Text garbled Line endings garbled Data in a PDF Data too granular Data entered by humans
 Data intermingled with formatting and annotations Aggregations computed on missing values
 Sample not random Margin-of-error too large Margin-of-error unknown Sample biased
 Data have been manually edited Inflation skews data Natural/seasonal variation skews data
 Timeframe manipulated Frame of reference manipulated Author untrustworthy
 Collection process opaque Data assert unrealistic precision Inexplicable outliers
 Data aggregated to wrong categories or geographies Results p-hacked Benford's Law fails
 An index masks underlying variation Data are in scanned documents Too good to be true”

⚠ **“DATA CLEANING” IS NOT A FORENSIC AUDIT**

Data cleaning programs correct logical inconsistencies, data duplications, impossible values, conflicting postal codes, outliers, other low-tide stuff. A well-designed cleaning program might identify 70% of Groskopf issues, but will have difficulties detecting systemic biases, falsification, and too-good-to-be true – which require experience and honest judgement. A virtue of AI forensic audits is their independent outsider status, unlike the ultimate insider Dr. Confirmation Bias.

★ **GET IT RIGHT FROM THE START IN BUILDING DATABASES: EARLY ADJUSTMENTS ARE CHEAP, LATER RESCUE REVISIONS COSTLY. EXPERT PRACTICAL ADVICE BY JEFF ATWOOD AND NATE SILVER**

“Of all the technical debt you can incur, the worst in my experience is bad names – for database columns, variables, functions, etc. Fix those IMMEDIATELY before they metastasize all over your code base and become extremely painful to fix later. . . and they always do.” JEFF ATWOOD
 “Going through some old data/code. One thing I've learned is when combining different datasets or doing complicated data processing, it pays to be compulsive about missing data or data that doesn't pass sanity checks. More often than you might think, the missing/miscoded/outlier cases indicate a larger, more systematic problem with your code or data. My advice is to do due diligence before moving onto next steps, because errors tend to compound.” NATE SILVER
 “Bioinformatics . . . or 'advanced file copying'” NICK LOMAN

SPREADSHEET AUDITS ARE ESSENTIAL

⚠️ PROVENANCE IS NOT DOCUMENTED

“Data are created by businesses, governments, nonprofits, nut-job conspiracy theorists. Data are gathered in different ways – surveys, sensors, satellites. Knowing where data came from provides huge insights into its limitations. Survey data is not exhaustive. Sensors vary in accuracy. Governments are disinclined to give you unbiased information. War-zone data are geographically biased due to dangers of crossing battle lines. Various sources are daisy-chained together. Policy analysts redistribute government data. Every stage in that chain is an opportunity for error. Know where your data came from.” Chris Groskopf

⚠️ BEWARE OF HUMAN-ENTERED DATA: THE CHIHUAHUA SYNDROME

“There is no worse way to screw up data than to let a single human type it in, without validation. I acquired a complete dog licensing database. Instead of requiring people registering their dog to choose a breed from a list, the system gave dog owners a text field to type into, so this database had 250 spellings of **Chihuahua**. Even the best tools can’t save messy data. Beware of human-entered data.” Chris Groskopf



⚠️ ELECTRONIC HEALTH RECORDS

⚠️ REDEFINING GROUPS MAY BREAK RANDOM ASSIGNMENT

Medical researchers sometimes redefine treatment groups to include only *patients treated* instead of *intended to treat*. But consider research randomly assigning patients to treatment vs. placebo: if the treatment takes place on the 5th floor in a building without elevators, for example, then less healthy patients may never arrive for treatment (example by Darrel Francis). Even if the treatment *has no effect at all*, the treated group will now appear to do better than placebo groups.

🚫 HOW DR. CONFIRMATION BIAS MESSES AROUND WITH MEASUREMENTS AND SPREADSHEETS

Effects in current medical research are small, large effects have already been discovered. And thus small pseudo-findings can be fabricated by a few cheats. Dr. Confirmation Bias doing data fudges: “Try 5 methods to manage missing data, see what works best.” “Patient compliance always averages out” “High/Low bins: try cuts at median, mean, midmean, isosceles harmonic mean” “Results not good, transform variables: log, arc sine, trichotomize, deca-chotomize, whatever it takes” “Report summary models only” “Look at absolute rates privately, report relative rates in press releases” “Too late to audit spreadsheet, findings optimized on unaudited data” “Consult with biostat folks at end of our data work, their job is to make our results truly significant: .001 or .01, **not** .05. After all, **our** research grant is paying the biostat guys. Whose side are they on?” “Find all subgroups where our desired findings work best, it’s only fair” “Stonewall all requests to see our original data” “If results contrary to our sponsor’s expectations: (1) Do not publish, (2) Stop it with randomized trials, (3) Actually, sponsor prefers no control arm at all, with outcomes measured by patient post-op/post-chemo Gratitude and Hope, never measure/report reductions in all-cause mortality.”

⚠️ BIOSTATISTICIANS DESCRIBE CHEAT REQUESTS MADE BY RESEARCHERS

522 consulting biostatisticians surveyed, and 75% responded. The survey reported these common inappropriate requests by researchers: “*removing or altering some data records to better support the research hypothesis; interpreting the statistical findings on the basis of expectation, not actual results; not reporting the presence of key missing data that might bias the results; ignoring violations of assumptions that would reverse the results.*”

“Researcher Requests or Inappropriate Analysis and Reporting: U.S. Survey of Consulting Biostatisticians,” Min Qi Wang, Alice F. Yan, Ralph V. Katz, *Annals of Internal Medicine*, 2018, edited

ASSESSING MEASUREMENT QUALITY

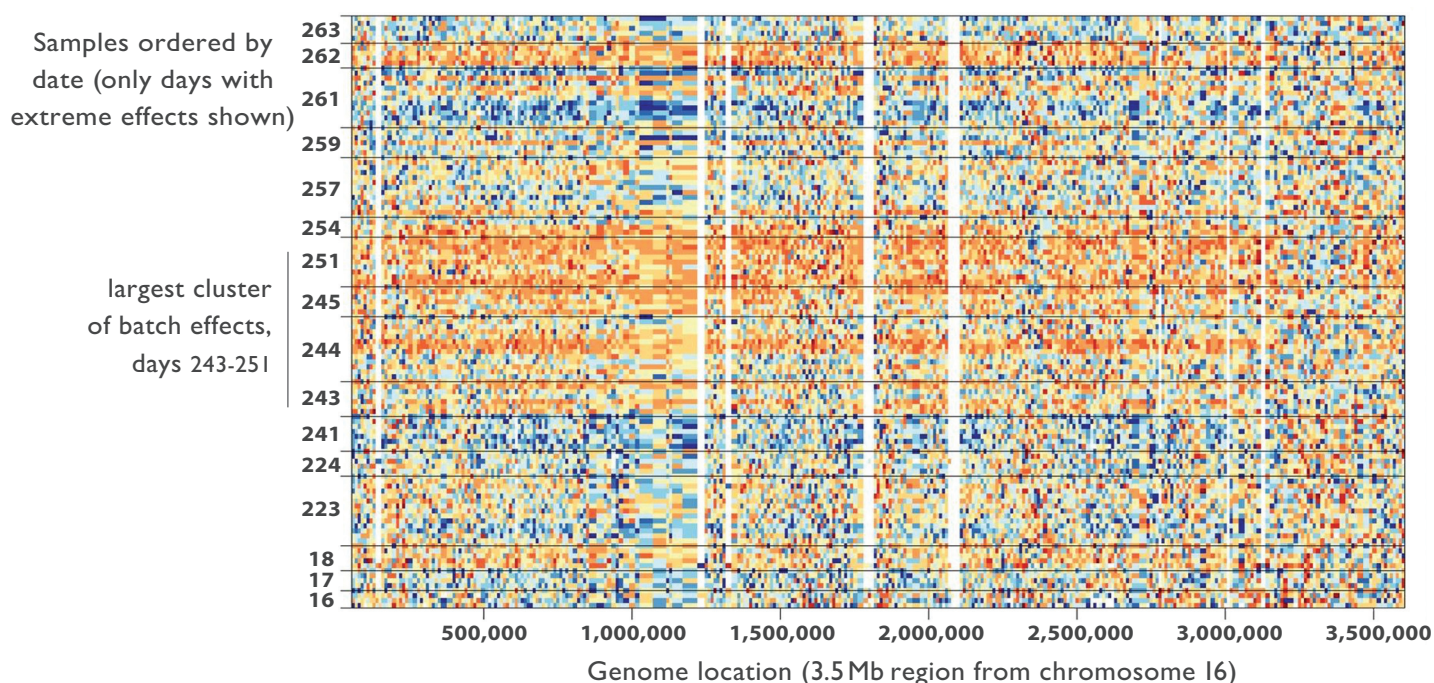
☆☆☆ DIRECTLY OBSERVE DATA COLLECTION AT THE EXACT MOMENT OF MEASUREMENT

☆☆☆ MEASUREMENT ERRORS ARE MEASURED BY OBSERVING DATA COLLECTION,
DETECTIVE WORK, DATABASE AUDITS – NOT BY STANDARD STATISTICAL ERROR ESTIMATIONS

⚠ BATCH EFFECTS IN SEQUENCING GENOME DATA

“High-throughput technologies are widely used to assay genetic variants, gene and protein expression, epigenetic modifications. Often overlooked complications are *batch effects*, occurring because measurements are affected by laboratory conditions, reagent lots, and personnel differences. This is a major problem when *batch effects are correlated with an outcome of interest and lead to incorrect conclusions*. Batch effects (and other technical and biological artifacts) are widespread and critical to address.”

Below, batch effects for 2nd-generation sequencing data from the 1000 Genomes Project. “Each row is a different HapMap sample processed in the same facility/platform. Samples are ordered by processing dates. Coverage data from each feature are standard across samples. Dark blue represents 3 standard deviations below average. Orange represents 3 standard deviations above average. Many batch effects are observed, and the largest one occurs between days 243–251 (long orange horizontal streaks).”



Jeffrey T. Leek, et al, “Tackling the widespread and critical impact of batch effects in high-throughput data,”
HHS Public Access authors manuscript, 2010, edited; *Nature Reviews Genetics* 11, 2010, 733-739.

⚠️ SURROGATE/PROXY/BIOMARKER MEASUREMENTS VS. MEDICAL PATIENT OUTCOMES

Measurements often serve as useful and convenient *surrogates/proxies/biomarkers* – which are based on *evidence/assumption/theory/folklore/marketing/convention* claiming that biomarkers signal actual patient outcomes. For example, markers such as blood pressure, lipids, patient hope and gratitude – rather than serious long-run outcomes such as incidence of strokes, heart attacks, quality of life, all-cause mortality. Only unbiased empirical evidence can assess links between markers and patient outcomes that are relevant, resilient, meaningful.

Changing hard outcomes to easier marker goals is a clear signal that the intervention doesn't work very well. For example, a new widely used oncology marker variable is *progression-free survival*, now playing a starring role in FDA Drug Approval Theater:

“Only one-third of cancer drugs entering Europe/U.S. markets have evidence of overall benefits in survival or quality of life. But regulators routinely approve new cancer drugs using surrogate endpoints measuring ‘progression-free survival,’ as the preferred surrogate endpoint in oncology drug trials, *not patient benefits*. After approval, drug companies have little or no incentive to evaluate the clinical benefit of their products. Data on overall survival or quality of life rarely emerge, even years after market entry. New cancer drugs should be approved on the basis of their overall survival and quality of life benefits.”

Christopher M. Booth,
Elizabeth A. Eisenhauer,
“Progression-Free Survival:
Meaningful or Simply
Measurable?,” *Journal of
Clinical Oncology*, 2012;
v.12, 10, edited

⚠️ “GROUND TRUTH” IS JUST SOMEONE ELSE’S SPREADSHEET

In ordinary language, “ground truth” is information collected on location, but database ground truthers never see their data at the moment of actual measurement. All such databases require independent forensic audits, checking for batch processing errors, changes and biases in measurement practices in space and time, and 100 other issues. Attributing truth to a database and claiming ‘proof of concept’ fools around with the meanings of truth and proof. This is the Fallacy of Equivocation: ‘whenever a term is used in 2 or more senses within a single argument, so that a conclusion appears to follow when it in fact does not,’ – as in the puns *ground truth, error, power, optimal, causal model, explained/unexplained variance, intellectual property*. Nature’s laws and survival rates are authentic Ground Truths.

⚠️ ERRORS IN, ERRORS OUT ⚠️ BIAS IN, BIAS OUT ⚠️ NO CAUSALITY IN, NO CAUSALITY OUT
⚠️ CIRCULARITY IN, CIRCULARITY OUT ⚠️ BIG DATA, ML, AI HUBRIS ⚠️ FALLACY OF EQUIVOCATION

⚠️ SURVIVAL BIAS

Most medieval castles were made of wood. We think most were made of stone because of survivor bias. Research databases are those that survived long enough to be selected for Ground Truth status. Survivor bias, subtle and inscrutable, requires deep meta-cognition and detective work about database provenance.



⚠️ IMMORTAL TIME BIAS ⚠️ SKETCHY IN, SKETCHY OUT ⚠️ BIAS IN, BIAS OUT

**⚠️ ENSHRINING AND REINFORCING PAST PRACTICES, A POLICE PREDICTION MODEL
SUFFERS FROM GARBAGE-IN/GARBAGE-OUT. BUT MODEL OUTPUTS, IN EFFECT,
MEASURE THEIR BIASES – AND SO PREDICTION MODELS DO RAT THEMSELVES OUT.**

Predictive policing is a widely used model. In daily work, some predictive systems may use “dirty data” that enshrines and intensifies past police practices. These models document the prevalence of unlawful and inefficient practices. There will never be better documentation, for it would be difficult to investigate/gather/model comparable data afresh. *Thus predictive data models directly measure prevailing biases!*

DIRTY DATA, BAD PREDICTIONS: HOW CIVIL RIGHTS VIOLATIONS IMPACT POLICE DATA,
PREDICTIVE POLICING SYSTEMS, AND JUSTICE. RASHIDA RICHARDSON, JASON SCHULTZ, KATE CRAWFORD

“Law enforcement agencies are increasingly using algorithmic predictive policing systems to forecast criminal activity and allocate police resources. Yet in numerous jurisdictions, these systems are built on data produced within the context of flawed, racially fraught, and sometimes unlawful practices (‘dirty policing’). This can include systemic data manipulation, falsifying police reports, unlawful use of force, planted evidence, unconstitutional searches. These policing practices shape the environment and the methodology by which data is created, which leads to inaccuracies, skews, and forms of systemic bias embedded in the data (‘dirty data’). Predictive policing systems informed by such data cannot escape the legacy of unlawful or biased policing practices that they are built on.

Nor do claims by predictive policing vendors that these systems provide greater objectivity, transparency, or accountability hold up. While some systems offer the ability to see algorithms used and even occasionally access to the data itself, there is no evidence to suggest that vendors independently or adequately assess the impact that unlawful and biased policing practices have on their systems, or otherwise assess how broader societal biases may affect their systems.

Confirmation Feedback Loops

Though there is research that empirically demonstrates that the mathematical models of predictive policing systems are susceptible to runaway feedback loops, where police are repeatedly sent back to the same neighborhoods regardless of the actual crime rate, such feedback loops are also a byproduct of the biased police data. More specifically, police data can be biased in two distinct ways. Fundamentally, police data reflects police practices and policies. If a group or geographic area is disproportionately targeted for unjustified police contacts and actions, this group or area will be over-represented in the data, in ways that often suggest greater criminality. Second, the data may omit essential information as a result of police practices and policies that overlook certain types of crimes and certain types of criminals. For instance, police departments, and predictive policing systems, have traditionally focused on violent, street, property, and quality of life crimes. Meanwhile, white collar crimes are comparatively under-investigated and over-looked in crime reports. Available studies estimate that 49% of businesses and 25% of households have been victims of white collar crimes, compared to a 1.1% prevalence rate for violent crimes and 7.4% prevalence for property crime.”

New York University Law Review Online February 13, 2019, edited.

**ELECTRONIC HEALTH RECORDS SEIZE OWNERSHIP OF MEDICAL PATIENT INFORMATION,
MEDICAL CENTER BUSINESS PLANS = OWN THE DATA, OWN THE PATIENT**

Medical patient ET, ending appointment: “Doctor, today I learned more about your EHR system than about my heart.”

Cardiologist, who failed to find recent test results, then ordered duplicate tests, says: “Me too.”

Electronic Health Records violate Tim Berners-Lee’s fundamental principle for data models:

“The hope is to allow a pool of information to develop, grow, and evolve. For this to be possible, **the method of storage must not place its own restraints on the information.**” Local EHRs, however, seize ownership by copyrighting patient content. Every day, logging into an EHR, millions of patients and staff must accept a bullying gag order that looks much like this parody:

Prior to proceeding, you must agree to every word below governing use of The System

All content included in the **Foundation University Healthcare Systems**, including, but not at all limited to, colors, words, photos, graphs, icons, buttons, graphics, images, videos, feature-length films, numbers (finite, rational, irrational, imaginary, real, troubled, prime), punctuation (including the Palatino Linotype interrobang), artworks, proper nouns, logos, trademarks, data (the “Content”), in all and any forms including compilations, are protected by all laws and conventions. Except as set forth no where, direct indirect reproduction (forget screenshots) of “Content” or The System, by any means, are prohibited without explicit consent of Interpol and **Foundation University Healthcare Systems**. **ACCEPT!** **DECLINE**

Don’t even think of clicking on **DECLINE** for you will arrive at The Shadowed Box threatening to **disable** your medical record – a mean nasty threat to frazzled patients. This is not a parody. Medical centers pitch empathy in their marketing, but intimidate patients to sign gag orders seizing ownership of all medical patient records.

If you fail to agree with **FUHS** Terms and Conditions, your medical record **will be disabled** and you will need to contact **Customer Services** to access your medical record.

GO BACK!

CONTINUE?

Patients have enough problems, and will give up on **Customer Services**. *Inconvenient opt-out* is inherent to software business models. Also medical patients appear to be redefined by the EHR as “users” and “customers.” Do medical patients thereby lose their unique legal rights?

In university medical centers, EHR systems may violate university norms: *freedom of speech and inquiry, civility and respect for others, even anti-plagiarism rules*. Gag orders also *stifle and block interface research* by prohibiting screenshots of any element of the EHR interface in research and professional meetings. Yet EHR interfaces do medical center command and control – and gag orders impede assessments by experts and researchers of this crucial interface. Why do university medical centers agree to this?

Proprietary EHRs are fundamental to U.S. medical center business models: own the data, own the patient, monetize everything, take over local competitors, make referrals to doctors within the System – and, in the U.S., charge monopoly prices, do predatory/surprise billing followed up by automated debt collectors that may bankrupt patients. Vendor capture of customers engages proven business models of proprietary software and Tony Soprano’s Waste Management. EHR Systems are governed by medical center suits, Senior Vice Presidents for Finance and Marketing. In the U.S., the suits maximize profits and report to commercial interests and investment bankers. The result is vast transfers of money from the sick to the rich. Financial interests directly conflict with the interests of patients, who seek to remain alive and healthy, and not bankrupted.

Electronic Health Records are *inherent to patient care*: the medical staff communicate with patients, adjust prescriptions, enter orders/notes, make referrals (inside to the System), and guide medical decisions. Despite vast public investment in EHRs, much of patient data is still communicated by fax or hand-carried by patients. Busy clinics receive thousands of fax pages daily. Hint to patients: get copies of all test reports, bring them to every medical appointment for you and your family.

In EHR encounters, are there grounds for a Medicare Quality of Care Grievance and a Plan Grievance Report (EHR rudeness)? Medicare rules for patient grievance reports (why no class-actions?) include:

"Examples of problems that are typically dealt with through the Quality of Care Grievance process: Duplicate tests, with possible side effects and adverse reactions. . ."

"Examples of problems that are typically dealt with through the Plan Grievance process: Disrespectful or rude behavior by doctors, nurses or other plan, clinic, or hospital staff. Long waiting times. Difficult to make appointments."

In medical care, everything within 50 meters of patients must follow fussy and detailed regulatory, industrial, professional standards. *Where is the evidence that EHRs are safe and effective, that benefits exceed harms to patients and medical staff?* To not answer these questions, each EHR installation has conducted one of the worst clinical trials ever: patients and staff are enrolled without consent in a vast unrandomized uncontrolled experiment, without measuring outcomes/harms/benefits, and with no plans for stopping the trial in event of excessive harms. And where was the Human Subjects Safety Review Board?

EHR problems are well-documented, and poignantly by Atul Gawande, *Why Doctors Hate Their Computers*. Eventually, perhaps, patients will some day own their medical records, despite intense resistance by U.S. hospital trade associations (which include University Medical Centers) seeking to own all patient records and to enforce their business models.

REMODELING MEDICAL PATIENT HEALTH RECORDS:

WHY A MEDICAL PATIENT'S HEALTH RECORD MUST BELONG TO THE PATIENT (BY ERIC TOPOL)

"It's your body	Your medical privacy is precious	You'd like to share it for medical research, but you can't get it	It requires comprehensive, continuous, seamless updating
You paid for it	The only way it can be made secure is to be decentralized	You have seen many providers in your life, but no health system/insurer has all your data	Your access or 'control' of your data is not adequate
It is worth more than any other type of data	It is legally owned by doctors and hospitals	No one (in US) has all their medical data from birth throughout their life	~10% of medical scans are unnecessarily duplicated due to inaccessibility of prior scans
It's widely sold, stolen, hacked. And you don't know it	Hospitals won't or can't share your data ('information blocking')	Your EHR was designed to maximize billing, not to help your health	You can handle the truth
It's full of mistakes, that keep getting copied and pasted, that you can't edit	Your doctor (>65%) won't give you copies of your office notes	You are more engaged, have better outcomes when you have your data	You need to own your data; it should be a civil right
You will be generating more of it, but it's homeless	You are far more apt to share your data than your doctor	Doctors with full access to patient records look at them routinely	It could save your life "

DATA MODELS CONTAIN BOTH REAL AND IMAGINARY PARTS

Millions of successful models thrive in physics, chemistry, engineering— because mathematical laws define ground truth. Human behavior research lacks such assurances: instead of Nature’s laws, we are stuck with after-the-fact statistical modeling. These models can produce handwaving, patent claims, disciplinary cults. How do we know if models are true and work? Models require experimental tests and applied interventions to learn the truth, just like real science:

‘Sailors talk about hydrodynamics the way CEOs talk about macroeconomics: they either treat it with mystical reverence, or they claim to understand it and are wrong. Unlike macroeconomics, though, if you know what you are doing you can test the propositions of hydrodynamics on actual physical models in a lab.’ BRENDAN GREELEY, FINANCIAL TIMES, MARCH 25, 2021

STATISTICAL MODELS: USES, ASSUMPTIONS, ADVERSE REACTIONS

Detailed package-inserts accompany prescription drugs. Consider a similar document for statistical models – describing use, pitfalls, prevalence of adverse effects, model breakdowns. *Imagine if encounters with every statistical model – in textbooks, computer code, workaday practice, publications – had to face up to adverse reactions. After all, these models can affect thousands of statistical lives.* At left, the first of 16 pages devoted to LISINOPRIL, a cardiovascular drug. At right, this mock-up document alerts users to model assumptions, breakdowns, adverse reactions – for a widely-used statistical model, stepwise logistic multiple regression

ADVERSE EFFECTS LISINOPRIL

The incidence of adverse effects varies according to the disease state of the patient:

People taking lisinopril for *treatment of hypertension* may experience the following side effects:

Headache (3.8%) Dizziness (3.5%) Cough (2.5%)
Difficulty swallowing or breathing (signs of angioedema)
allergic reaction (anaphylaxis)
Hyperkalemia (2.2% in adult clinical trials)
Fatigue (1% or more) Diarrhea (1% or more)
Some severe skin reactions have been reported rarely,
including toxic epidermal necrolysis and Stevens-Johnson
syndrome; causal relationship has not been established.

People taking lisinopril for treatment of *myocardial infarction* may experience the following side effects:

Hypotension (5.3%) Renal dysfunction (1.3%)

People taking lisinopril for the treatment of *heart failure* may experience the following side effects:

Dizziness (12% at low dose – 19% at high dose)
Hypotension (3.8%) Chest pain (2.1%)
Fainting (5-7%)
Hyperkalemia (3.5% at low dose – 6.4% at high dose)
Difficulty swallowing or breathing (signs of angioedema),
allergic reaction (anaphylaxis)
Fatigue (1% or more) Diarrhea (1% or more)
Some severe skin reactions (toxic epidermal necrolysis,
Stevens-Johnson syndrome) have been reported rarely,
causal relationship not established.

BLACK BOX WARNING !

STEPWISE LOGISTIC MULTIPLE REGRESSION:
DO NOT USE FOR MAKING CAUSAL INFERENCES.
APPROVED ICU USE: COMPASSIONATE THERAPY
FOR STAGE 3 DUSTBOWL EMPIRICISM

ADVERSE EFFECTS

STEPWISE LOGISTIC MULTIPLE REGRESSION (SLMR)
(TRADE NAMES: MACHINE-LEARNING, ML, AI)

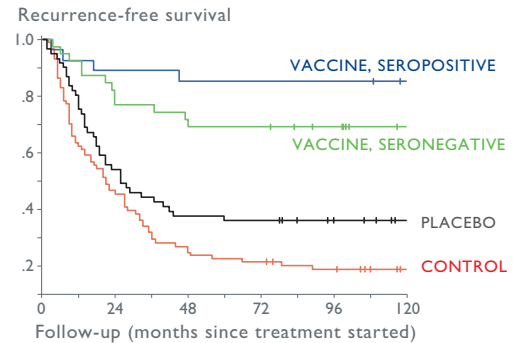
SLMR-ML data analysts may experience adverse outcomes:
Cherry-picked models (80.1%) Model and p-hacking (>70%)
Ioannidis syndrome: most published medical research is false
Unpublishable by highly-ranked medical journals (40.0%)
Dizziness (23%) Fatigue (11.8% or more) Fainting (7.3%)
Table 2 fallacy: confounding direct and indirect causes (100%)
Multicollinearity (80.5%) Hankins Condition: ‘A journey of
a thousand hypotheses begins with a single SLMR’ Failure to
replicate (70.0%) Accusations of garbage in, garbage out.

Model assumptions: one-way independent causes, errors
independently and identically distributed, x -variables
measured without error. Feedback, simultaneity, interaction
effects assumed not to exist; or, if modeled, more assumptions
and more data are needed. Implicitly assumes SLMR finds
the single best equation, practical experience suggests there
are better models with different variables.

Daniel Westreich and Sander Greenland, “The Table 2 Fallacy,”
Am J Epidemiol, 2013, 177, 4, 292–298; Cuthbert Daniel and
Fred Wood, *Fitting Equations to Data*, 1980, 84–85; Gary Smith,
“Step away from stepwise,” *Big Data*, 2018, 5:32.

☆☆☆ EXEMPLARY REPORT OF POSSIBLE SIDE-EFFECTS OF KAPLAN-MEIER SURVIVAL ESTIMATES

Kaplan-Meier curves track survival times, numbers of patients living over a period of time after a medical intervention. The abstract of this famous paper (>50,000 citations) warns that “lifetime (age at death) is independent of potential loss time; in practice this assumption deserves careful scrutiny.” KM lines show data directly, and the quality of inference depends on the character of the data. Engineers at JnF Practical Quality Control have produced an excellent package insert on issues in KM survival times. *Each use of a model should remind users of constraints, assumptions, breakdowns.*



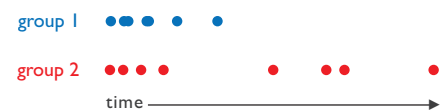
“Implicit factors Lack of independence within a sample is often caused by an implicit factor in the data. For example, if we are measuring survival times for cancer patients, diet may be correlated with survival times. If we do not collect data on implicit factors (diet in this case), and the implicit factor has an effect on survival times, then we in effect no longer have a sample from a single population, but instead a sample that is a mixture drawn from several populations, one for each level of the implicit factor, each with a different survival distribution. Implicit factors affect censoring times, by affecting the probability that a subject will be withdrawn from the study or lost to follow-up. For example, younger subjects may move away (and be lost to follow-up) more frequently than older subjects, so that age (an implicit factor) is correlated with censoring. If the sample under study contains many younger people, the results of a study may be substantially biased because of different patterns of censoring. This violates the assumption that censored values and noncensored values all come from the same survival distribution. Stratification can control for an implicit factor.

Lack of independence of censoring If pattern of censoring is not independent of survival times, then survival estimates may be too high (if subjects who are more ill tend to be withdrawn from the study), or too low (if subjects who will survive longer tend to drop out of the study, lost to follow-up). The estimates for the survival functions and their variances rely on independence between censoring times and survival times. If independence does not hold, the estimates may be biased, and the variance estimates may be inaccurate. An implicit factor not accounted for by stratification may lead to a lack of independence between censoring times and observed survival times.

Lack of uniformity within a time interval Kaplan-Meier estimates for survival functions and standard errors rely on assumptions that the probability of survival is constant within each interval (although it may change from interval to interval), where the interval is the time between two successive noncensored survival times. If the survival rate changes during the course of an interval, then the survival estimates for that interval will not be reliable or informative.

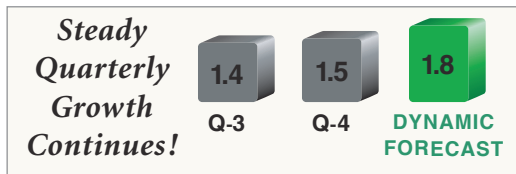
Many censored values A study may end up with many censored values, from having large numbers of subjects withdrawn or lost to follow-up, or from having the study end while many subjects are still alive. Large numbers of censored values decrease the equivalent number of subjects exposed (at risk) at later times, making Kaplan-Meier estimates less reliable than they would be for the same number of subjects with less censoring. Moreover, if there is heavy censoring, the survival estimates may be biased (because the assumption that all censored survival times occur immediately after their censoring times may not be appropriate), and estimated variances become poorer approximations, perhaps considerably smaller than the actual variances. A high censoring rate may also indicate problems with the study: ending too soon (many subjects still alive at the end of the study), or a pattern in censoring (many subjects withdrawn at the same time, younger patients lost to follow-up sooner than older ones). If the last observation is censored, the Kaplan-Meier estimate of survival can not reach 0.

Patterns in plots of data If the assumptions for the censoring and survival distributions are correct, then a plot of either the censored or the noncensored values (or both together) against time should show no particular patterns, and the patterns should be similar across the various groups.

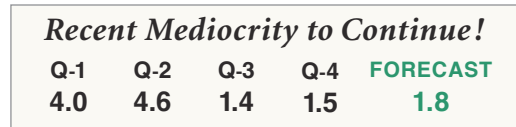


Special problems with small sample sizes Time intervals in Kaplan-Meier calculations are determined by distinct noncensored survival times. This means that the smaller the sample size, the longer the intervals will be, raising questions of whether the assumption of a constant survival probability within each interval is appropriate. Small samples make it difficult to detect possible dependencies between censoring and survival, or the presence of implicit factors. If the number of subjects exposed (at risk) in an interval or the number of subjects that survived to the beginning of that interval is small, variance estimates for survival functions will tend to underestimate actual variance. This situation is most likely to occur for later intervals, when most subjects have either died or been censored, so that variance estimates for later intervals are less reliable than those for earlier intervals.”

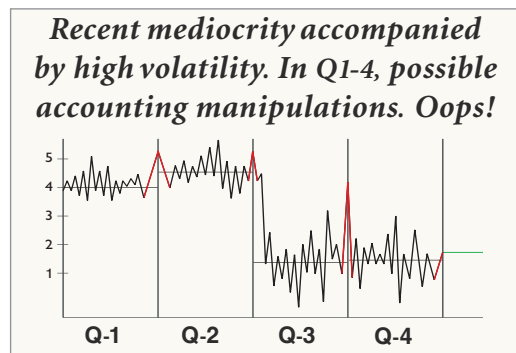
GLIB, CHERRY-PICKED, BIASED SUMMARIES AVOID SHOWING THE DATA



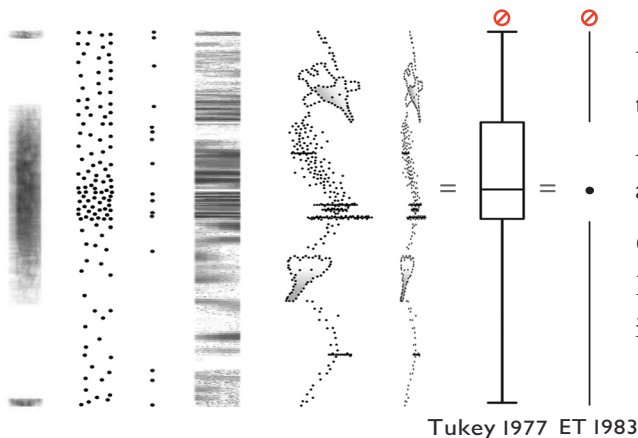
Summary reports are economical with the truth: this slide cherry-picks, over-summarizes the data – and disrespects the audience. “The Board of Directors loves good news. Our internal numbers are not entirely good news. Need to fix it in our slides. Everyone does it.”



Two additional quarters of data change the story. How about 12 more quarters? But regardless of how many quarters, readers are shown *binned quarterly data*, just one number per quarter. LittleDataGraphics are friends of falsity, enemies of truth.



Detailed data moves closer to the truth. No more binning, less cherry-picking, less truncation. This graphic reveals increases in recent volatility. **End-of-quarter-upticks** may signal accounting manipulation (premature revenue recognition). This time-series is easily readable by all – Boards of Directors, financial journalists, shareholders, colleagues. The initial slide    treats viewers like mushrooms    kept in the dark in manure.



Upon learning about averages, every schoolchild knows that divergent data sets can produce identical summaries. At left, 6 different data sets and their 63 possible combinations all yield the same identical boxplot.

Censoring data often produces false findings. Show the data. Nowadays display screens have enormous resolution. The 2021 iPhone screen shows 3,566,952 pixels, 458 per square inch.

Can boxplots chase down outliers, as *Computational Methods in Physics* (2018) claims? Why show only the 2 extreme min/max outliers and bin all the other data? Unlike [this LittleData boxplot](#), an unbinned plot of the same data shows all 1000 measurements:

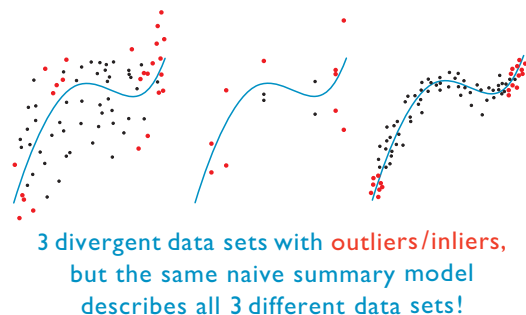


Now the 2-dimensional case:



A summary model conceals the data and inliers and outliers

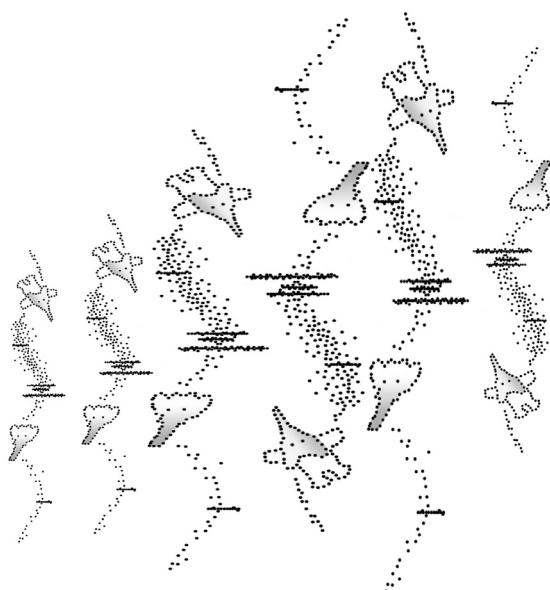
like boxplots, glib published summary models lack credibility



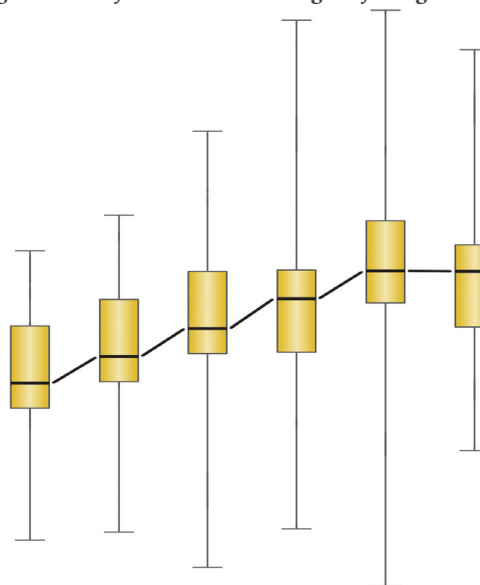
BINNING CONTINUOUS DATA IS MEDIUM-QUALITY EVIDENCE OF FALSIFICATION

Multiple boxplots cheat. Golden boxplots double-bin the flying elephants data set: X into 6 columns, Y into quartiles/outliers – making false inferences. Binning creates thresholds/plateaus notoriously difficult to replicate or verify in research on humans. Thresholds are model-imposed, not data-driven. Some medical research reports are written by the sponsor's ghostwriters who spin data into a product pitch. Sponsors can suppress the publishing of research contrary to desired results. Publication bias corrupts the collective knowledge (meta-analysis) of entire research fields.

ORIGINAL XY-DATA
12 elephants flying in sinusoidal
flight-paths with identical local pulses. $N = 2,598$



BOXPLOTS = FALSE FINDINGS
"As X dosage increases, response Y makes steady gains to reach a plateau – a highly significant and novel threshold – that signals need for increased dosages of drug X ." $N = 2,598$



⊗ BOXPLOT CHEATS ⊗ DOUBLE-BINNING ⊗ IMAGINARY THRESHOLDS/PLATEAUS/CUT-POINTS ⊗ SPINNING
⊗ GHOSTWRITING, GHOST DATAVIZ ⊗ PUBLICATION BIAS ★ FOLLOW THE MONEY ★ SHOW THE DATA

DATA GRAPHICS AND DATA AVAILABILITY: HOW DR.CONFIRMATION BIAS JUSTIFIES CHEATING

"Smoothed data summaries reduce clutter, make our results understandable to journalists/doctors/sponsors. Readers don't want *data data data*! They love simple graphics with a strong message. This isn't rocket science. Frankly, readers look only at abstracts, graphics, citations – so that's where our team works hard to pitch our findings. Reasons for not making our data available: Trade secret. Violate patient privacy. Hard drive crashed. Intellectual property. In litigation. Double top secret. Patent pending. IPO silent period. All of the above."

REPLY TO DR.CONFIRMATION BIAS: STOP CHEATING

"Clutter" in data graphics is evidence that your models don't fit the data – and that you know it. You also know that your summary graphics cover up contrary data and depict dubious thresholds not present in the data set. Such cheats are obvious and easily detected, and damage your credibility.

On high-resolution data: every day a billion people look at e-maps with data densities 20 times greater than your deceptive LittleDataGraphics. Mapmakers and scientists publish readable graphics showing immense data. Why assume that readers suddenly become stupid just because they're reading your research report? To improve learning from data, credibility, and integrity, show the data.

HOW TO EVALUATE AND REPORT THE CREDIBILITY OF RESEARCH

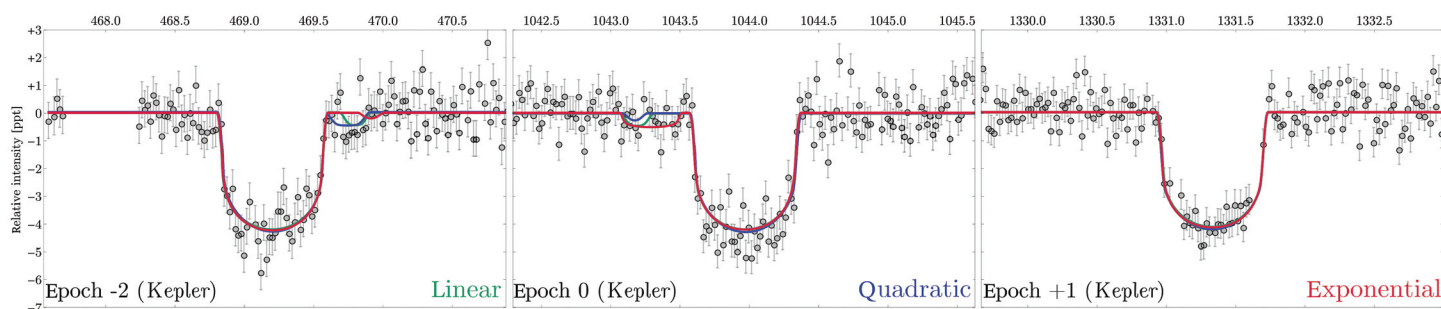
Evidence for a large exomoon orbiting Kepler-1625b

Alex Teachey and David M. Kipping *Science Advances* 4,10, 03 October 2018, edited.

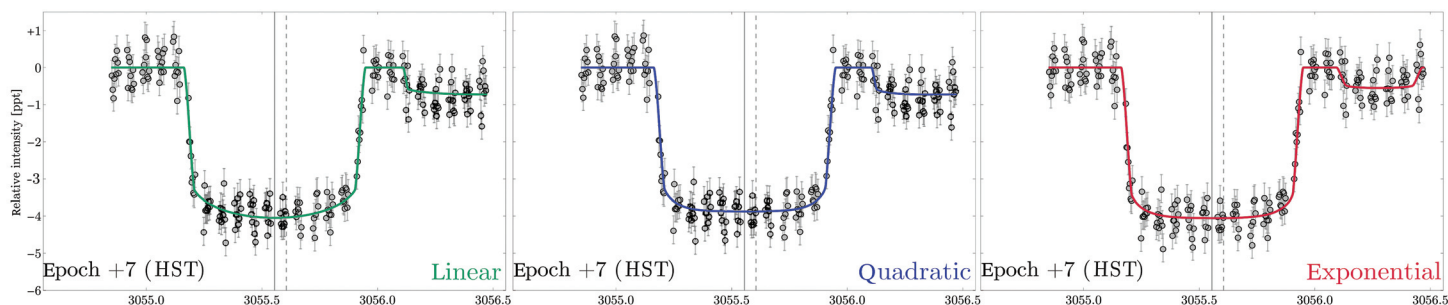
“Exomoons are the natural satellites of planets orbiting stars outside our solar system. We present new observations of a candidate exomoon associated with Kepler-1625b using the Hubble Space Telescope to validate or refute the moon’s presence. We find evidence in favor of the moon hypothesis, based on timing deviations and a flux decrement from the star consistent with a large transiting exomoon. Self-consistent photodynamical modeling suggests that the planet is likely several Jupiter masses, while the exomoon has a mass and radius similar to Neptune. Since our inference is dominated by a single but highly precise Hubble epoch, we advocate for future monitoring of the system to check model predictions and confirm repetition of the moon-like signal.”

This paper exemplifies the authentic presentation of data and uncertainties. The summary is written in straight-forward language describing the findings and need for replication. Unlike many medical research publications, this paper is not paywalled; computer code and curve-fitting are open-source and replicable, not absurdly claimed to be intellectual property, and the exomoon is not patented, trademarked, lawyered up, monetized.

These exoplanet data graphics show thousands of measurements along with 3 competing models (linear, quadratic, exponential) from 2 data sources—with the resolution and readability worthy of Google maps. These data displays show sensitivity tests of the statistical models and their data sources:



Moon solutions. The three transits in Kepler (top) and the October 2017 transit observed with HST (bottom) for the three trend model solutions. The three colored lines show the corresponding trend model solutions for model M, our favored transit model. The shape of the HST transit differs from that of the Kepler transits owing to limb darkening differences between the bandpasses.



Alex Teachey and David M. Kipping, “Evidence for a large exomoon orbiting Kepler-1625b”
Science Advances 4, 10, 03 October 2018

MAKING JUDGEMENTS ABOUT UNCERTAINTY, CREDIBILITY, TRUTH IN EMPIRICAL RESEARCH

Uncertainties are inherent in data analysis. Uncertainties and errors are entangled with one another and just about anything else that moves – contrary to standard statistical models, where a few over-modeled numbers based on empirically false assumptions create an illusion of certainty. Rather, the credibility of evidence-based conclusions requires detective work about measurements, how those measurements were analyzed, the character and quality of the substantive explanatory theory and its competitors, and avoiding pitfalls. The exomoon research report concludes with **exemplary judgements of uncertainty**. It is certain that exoplanets exist (~5000 already identified), and moons accompany planets (~200 moons in our solar system). But is the model describing a *real* exomoon or creating an *illusory* exomoon – perhaps due to anomalies in signal processing, or perhaps it is an *exoplanet*, not an *exomoon*?

Together, a detailed investigation of a suite of models tested in this work suggests that the exomoon hypothesis is the best explanation for the available observations. The two main pieces of information driving this result are (i) a strong case for TTVs [transit timing variations] in particular a 77.8-min early transit observed during our HST observations, and (ii) a moon-like transit signature occurring after the planetary transit. We also note that we find a modestly improved evidence when including additional dynamical effects induced by moons aside from TTVs.

3 counter-explanations are countered by specific evidence. Are there other counter-explanations, known or unknown?

The exomoon hypothesis is further strengthened by our analysis that demonstrates that (i) the moon-like transit is not due to an instrumental common mode, residual pixel sensitivity variations, or chromatic systematics; (ii) the moon-like transit occurs at the correct phase position to also explain the observed TTV; and (iii) simultaneous detrending and photodynamical modeling retrieves a solution that is not only favored by the data but is also physically self-consistent.

A thoughtful sentence, then followed up by acknowledging three remaining issues.

Together, these lines of evidence all support the hypothesis of an exomoon orbiting Kepler-1625b. The exomoon is also the simplest hypothesis to explain both TTV and post-transit flux decrease, since other solutions would require two separate and unconnected explanations for these two observations. There remain some aspects of our present interpretation of the data that give us pause. First, the moon's Neptunian size and inclined orbit are peculiar, though it is difficult to assess how likely this is *a priori* since no previously known exomoons exist. Second, the moon's transit occurs toward the end of the observations and more out-of-transit data could have more cleanly resolved this signal. Third, the moon's inferred properties are sensitive to the model used in correcting Hubble Space Telescope's visit-long trend, and thus some uncertainty remains regarding the true system properties. However, the solution we deem most likely, a linear visit-long trend, also represents the most widely agreed upon solution for the visit-long trend in the literature.

Formal models yield a goes-to-eleven Bayes factor, but then followed up by list of concerns that can't be quantified.

All in all, it is difficult to assign a precise probability to the reality of Kepler-1625b-i [the possible exomoon]. Formally, the preference for the moon model over the planet-only model is very high, with a Bayes factor exceeding 400,000. But, this is a complicated and involved analysis where a minor effect unaccounted for, or an anomalous artifact, could potentially change our interpretation. In short, it is the unknown unknowns that we cannot quantify. These reservations exist because this would be a first-of-its-kind detection—the first exomoon. Historically, the first exoplanet claims faced great skepticism because there was simply no precedence for them. If many more exomoons are detected in the coming years with similar properties to Kepler-1625b-i exomoon, it would hardly be a controversial claim to add one more.”

In reasoning about uncertainty, *known* and *unknown unknowns* are surely the case, and are acknowledged here, and should be similarly acknowledged in medical research publications.

Alex Teachey and David M. Kipping, “Evidence for a large exomoon orbiting Kepler-1625b,” *Science Advances* 4, 10, 03 October 2018, edited.

FAILURE AND FALSITY OF CLASSICAL STATISTICAL MODELS OF UNCERTAINTY

Flying at high altitude over a crime scene, standard statistical analysis is based on puns – ‘error’ ‘confidence’ ‘unexplained variance’ ‘significant’ ‘power’ ‘independent’ – *but empirical uncertainties and errors are not detected/measured/modelled by standard statistical methods*. It is falsification to estimate ‘uncertainties’ under false assumptions, announce a hypothesis is ‘true’ or a numerical difference is ‘significant’. Standard model assumptions, necessary for mathematical tractability, are briefly described in textbooks, then forgotten/hidden in workaday data analysis and published reports.

Are assumption-bound classical error models too certain and exact about too many things – and thus too certain and exact for evaluating empirical uncertainty? Links between math models and noisy reality are created by punning, calling two different things by the same name. Models also include information about the attitudes of researchers and even readers, although the data doesn’t care. These ambiguities/puns produced 100 years of insider epistemological debates about meaning of statistical credibility among readers, researchers, journal editors, statistical societies. Leave epistemology to the Departments of Philosophy and History of Science.

CREDIBLE DATA-BASED CONCLUSIONS

In Pandemic month 5, investigator-initiated research by the Oxford Randomized Evaluation of Covid-19 Group found dexamethasone reduced mortality among critically-ill patients:

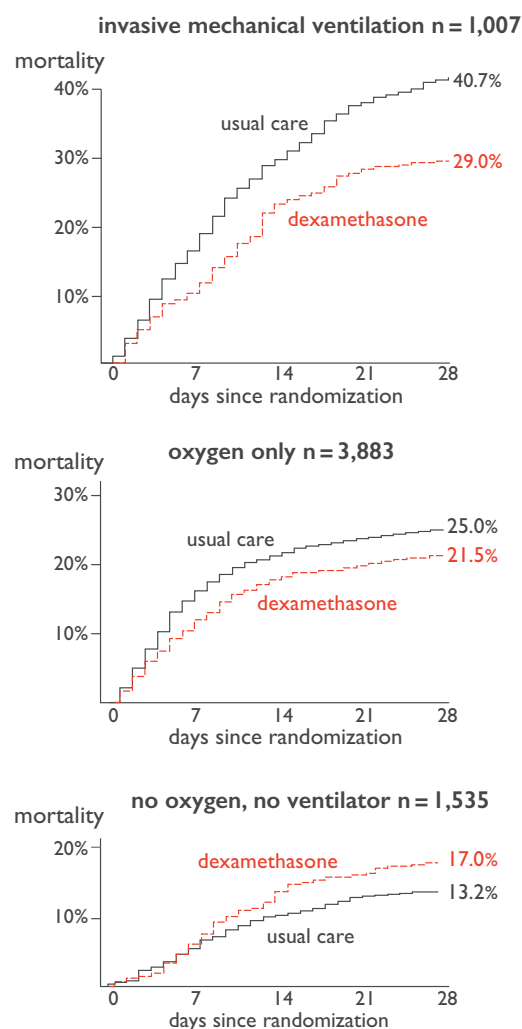
‘In patients hospitalized with COVID-19, dexamethasone reduced 28-day mortality among those receiving invasive mechanical ventilation or oxygen at randomization, but not among patients not receiving respiratory support. Dexamethasone reduced deaths by one-third in patients receiving invasive mechanical ventilation (29.0% vs 40.7%, RR 0.65), and by one-fifth in patients receiving oxygen without invasive mechanical ventilation (21.5% vs 25.0%, RR 0.80), but did not reduce mortality in patients not receiving respiratory support at randomization (17.0% vs 13.2%, RR 1.22).’

‘For less than £50 (US\$63), you can treat eight patients and save one life.’ Dexamethasone is already available worldwide. (Given an actual cost of \$63, U.S. hospitals will likely try to collect \$800 to \$5,000.)

This study is timely and smart: a randomized controlled trial, many UK sites, *all-cause mortality* is the measured outcome (not proxies, surrogates, markers), a sufficient *n*, relative/absolute risks shown together. This rapid RCT platform then assessed many other covid-19 interventions.

Few financial conflicts for these researchers: they do public health, not profits. Covid-19 treatments are difficult enough, without limiting research solely to new patented drugs.

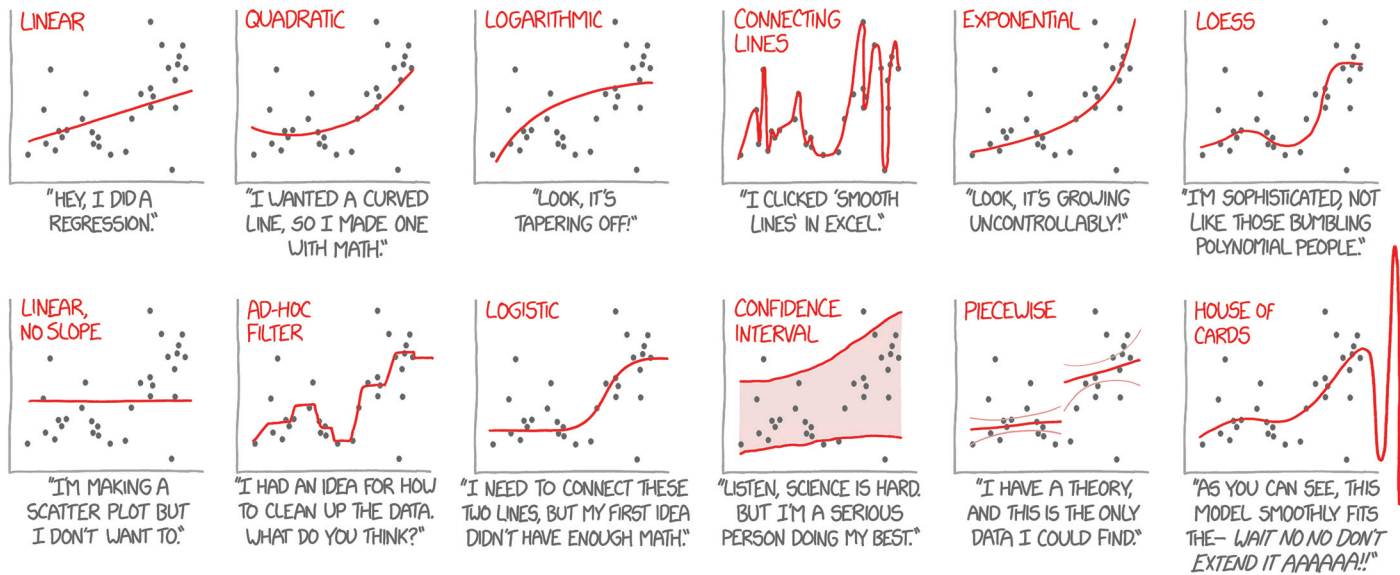
Often first-discovery evidence is the most enthusiastic that will ever be found, *too good to be true*. Independent replications are a necessity. Ten weeks later, the World Health Organization published meta-analyses of 6 small new RCTs + the original study, now with data from 12 countries – and replicated the original findings, and extended the results to other corticosteroids.



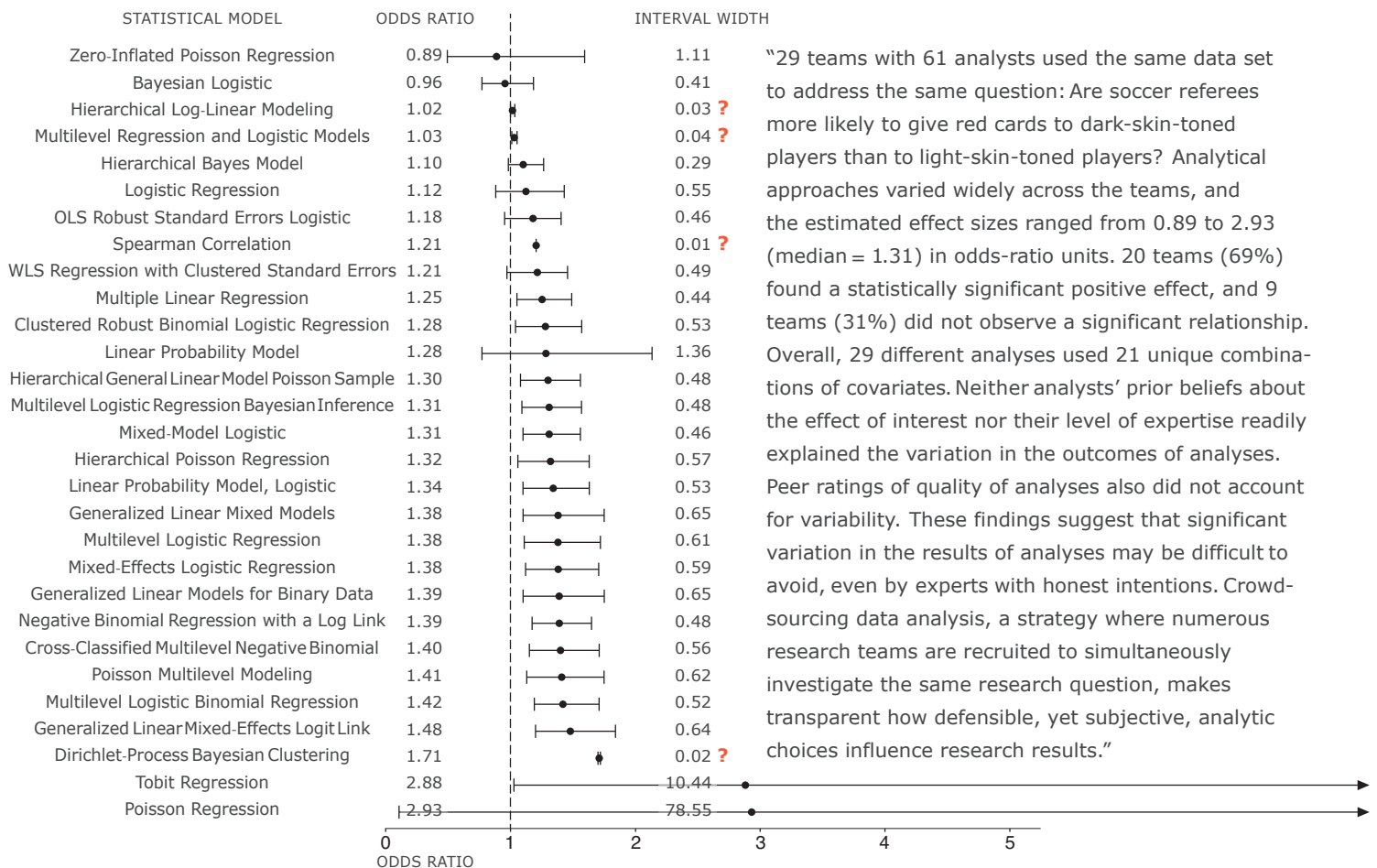
Peter W Horby, Martin J Landray, et al, ‘Effect of dexamethasone in hospitalized patients with covid-19,’ June 22, 2020, online preprint, edited; Kai Kuperschmidt, ‘One U.K. trial is transforming COVID-19 treatment. Why haven’t others delivered more results?’ *Science*, July 2, 2020. ‘Association Between Administration of Systemic Corticosteroids and Mortality Among Critically Ill Patients With COVID-19,’ meta-analysis by WHO Rapid Evidence Appraisal for COVID-19 Therapies (REACT) Working Group, *JAMA*, published online September 2, 2020.

MODEL MULTIPLICITY: SAME DATA, BUT DIFFERENT MODELS, MODELERS, MESSAGES

xkcd CURVE-FITTING METHODS AND THE MESSAGES THEY SEND



USING 21 DIFFERENT SETS OF EXPLANATORY VARIABLES AND MANY ARCAINE MODELS, 29 DATA ANALYSIS TEAMS PRODUCED MAYBE SOMEWHAT CONSISTENT FINDINGS FROM THE SAME SMALL DATA SET



DATA MODELING: INHERENT ISSUES, BUT FEW KNOWN PREVALENCE RATES

⚠ MODELS WITH PLENTIFUL PARAMETERS CAN FIT PROTONS, ELEPHANTS, NOISE, WHATEVER

No matter how big one's proton detector, ever more extravagant Grand Unified Theory models can always be constructed that elude tests – such as symmetry groups E_6 or E_8 , whose plentiful parameters can be tuned to make protons live as long as one pleases. One model might be correct, but no one would ever know. Dimitri Nanopoulos said: “People can construct models with higher symmetries and stand on their nose and try to avoid proton decay.

OK, you can do it, but you cannot show it to your mother with a straight face.” NATALIE WOLCHOVER

With four parameters I can fit an elephant, with five I can make it wiggle its trunk. JOHN VON NEUMANN

⚠ MODEL MULTIPLICITY AND CURVE-FITTING: WILLIAM FELLER'S ADVERSE REACTION

An unbelievably large literature tried to establish a transcendental “law of logistic growth.” Lengthy tables, complete with chi-square tests, supported this thesis for human populations, bacterial colonies, development of railroads, etc. Both height and weight of plants and animals were found to follow the logistic law even though it is theoretically clear these 2 variables cannot be subject to the same distribution. ^{height is linear, weight is volume} The trouble with the theory is that not only the logistic distribution, but also the normal, the Cauchy, and other distributions can be fitted to the same material with the same or better goodness of fit. In this competition logistic distributions play no distinguished role whatever; many theoretical models can be supported by the same observational material. Theories of this nature are short-lived because they open no new ways, and new confirmations of the same old thing soon grow boring. But the naive reasoning has not been superseded by common sense. WILLIAM FELLER

⚠ PEOPLE CAN'T KEEP THEIR OWN SCORE:

CLAIMING “MY DOG IS THE BEST DOG IN THE WORLD” DOES NOT VALIDATE YOUR MODEL

If someone shows you simulations that only show the superiority of their method, you should be suspicious. Good simulations will show where the method shines but also where it breaks. BYRAN SMUCKER & ROB TIBSHIRANI

Tuning your own method but insufficiently tuning the competing methods is one of those hidden problems in simulations for methods papers. MANJARI NARAYAN & ROB TIBSHIRANI

⚠ A GOOD MODEL EXPLAINS DATA, DOES NOT MEMORIZE DATA

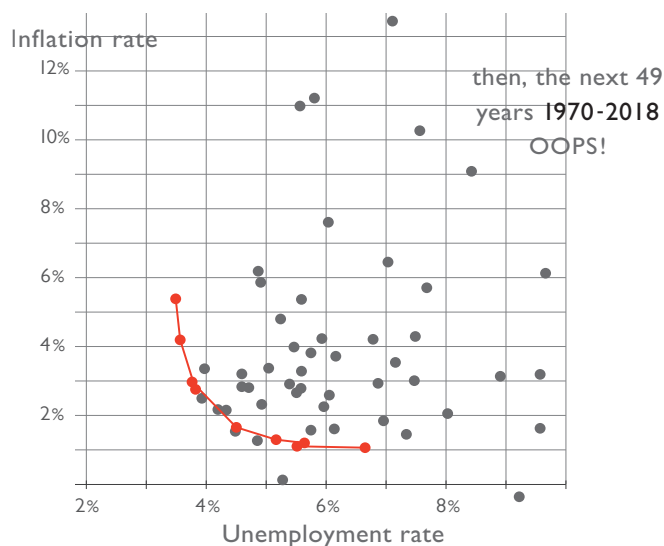
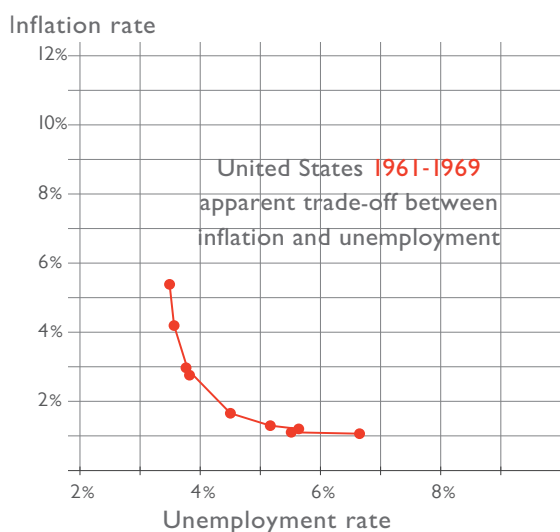
A mesh to mimic content makes luscious animations, but mesh-drapery doesn't explain much. Overfitted models chasing data are brittle, breaking down and regressing toward the truth when confronted by new data. Computing millions of models is easy, but explaining something so well that it leads to replicated real-world explanations and successful interventions is very difficult.



In that Empire, the Art of Cartography attained such Perfection that the map of a single Province occupied the entirety of a City, and the map of the Empire, the entirety of a Province. In time, those Unconscionable Maps no longer satisfied, and the Cartographers Guilds struck a Map of the Empire whose size was that of the Empire, and which coincided point for point with it. The following Generations, who were not so fond of the Study of Cartography as their Forebears had been, saw that this vast Map was Useless, and not without some Pitilessness, they delivered it up to the Inclemencies of Sun and Winters. In the Deserts of the West, still today, there are Tattered Ruins of that Map; in all the Land there is no other Relic of the Disciplines of Geography. JORGE LUIS BORGES

⚠️ FRESH DATA REMODELS MODELS: ON THE PHILLIPS CURVE,

A FOUNDATIONAL MODEL IN MACROECONOMIC RESEARCH, TEXTBOOKS, AND POLICY-MAKING



⚠️ THE CURSES (AND BLESSINGS) OF HIGH-DIMENSIONAL DATA

“The curse of dimensionality arises when analyzing data in high dimensional spaces (often with hundreds or thousands of dimensions) that do not occur in low-dimensional settings such as the 3-dimensional physical space of everyday experience. The expression was coined by Richard Bellman when considering problems in dynamic optimization. Cursed phenomena occur in domains such as numerical analysis, sampling, combinatorics, machine learning, data mining, databases. When the dimensionality increases, the volume of the space increases so fast that the available data become sparse. In order to obtain a statistically sound and reliable result, the amount of data needed to support the result often grows exponentially with the dimensionality. Also, organizing and searching data often relies on detecting areas where objects form groups with similar properties; in high-dimensional data, however, all objects appear to be sparse and dissimilar in many ways.” WIKIPEDIA

The curse is a mathematical truth, but it's more complicated than that. Advanced students may consult David Donoho, “High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality,” 2000.

⚠️ SUBGROUP ANALYSIS: LEARNING FROM DATA VS. SUB-SUB-SUB-GROUP CHERRY-PICKING

In learning from data, subgroup analysis is essential. But, in workaday practice, a big problem: cherry-picking subgroups for desired findings, then publishing those findings as proven results, rather than as possibilities requiring replication. John Mandrola reported an extreme example of subgroup over-reach: “The clinical trial compared percutaneous transluminal coronary angioplasty (PTCA) against standard medical therapy for angina. The primary endpoint of death and myocardial infarction occurred in 6.3% of patients in the PTCA arm and only 3.3% in the medical arm.

But instead of saying angioplasty was twice as bad as medical therapy, the abstract begins with this:

In patients with coronary artery disease considered suitable for either

PTCA or medical care, early intervention with PTCA [subgroup]

was associated with greater symptomatic improvement [secondary endpoint subsubgroup],

especially patients with more severe angina [secondary endpoint subsubsubgroup].”

**MANY MEDICAL INTERVENTIONS ARE UNVALIDATED, INVALIDATED, LOW VALUE:
IS NON-EVIDENCE BASED MEDICINE MORE PREVALENT THAN EVIDENCE-BASED MEDICINE?**

“An analysis of 3,017 randomized controlled trials published in 3 leading medical journals (*Journal of the American Medical Association*, *The Lancet*, *New England Journal of Medicine*) identified 396 medical reversals. Low-value medical practices are either ineffective or cost more than other options but only offer similar effectiveness. Such practices can result in physical and emotional harms, undermine public trust in medicine, and have opportunity and financial costs. Identifying and eliminating low-value medical practices will reduce costs and improve care. Medical reversals are a subset of low-value medical practices and are defined as practices that have been found, through randomized controlled trials, to be no better than a prior or lesser standard of care.”

Diana Herrera-Perez, Alyson Haslam, Tyler Crain, Jennifer Gill, Catherine Livingston, Victoria Kaestner, Michael Hayes, Dan Morgan, Adam S. Cifu, and Vinay Prasad, “A Comprehensive Review of 3000 Randomized Clinical Trials in 3 Leading Medical Journals Reveals 396 Medical Reversals,” *eLife*. 2019; 8: e45183. published online 2019 June 11.

A short list of medical reversals

low dose aspirin for primary prevention of cardiovascular events
surgery for meniscal tear or knee arthritis (~1,000,000 surgeries yearly) magnesium supplements for leg cramps
vitamin pills to improve health multivitamins for prevention of cardiovascular disease
wearable tech for long term weight loss avoidance of peanut allergy by infant peanut exposure
tight blood sugar control in critically ill patients bedrest to prevent preterm birth
mammogram screening every 2 years mammogram screening for all women
MRI in breast cancer surgery compression stockings to reduce risk of deep vein thrombosis after stroke
intravenous drug administration during out-of-hospital cardiac arrest
epidural glucocorticoid injections for spinal stenosis (> \$200 million annually)
carotid artery stenting (compared to surgery) for symptomatic carotid stenosis
Ginkgo biloba for preventing cognitive decline in older adults (\$250 million annually)
vitamin E in primary prevention of cardiovascular disease
electrocardiographic/hemodynamic effects of diet supplements containing Ephedra
opioid-based analgesics for acute extremity pain in the emergency department
cardiovascular effects of intensive lifestyle intervention in type 2 diabetes
screening tests and all-cause mortality HRT for preventing chronic disease in post-menopausal women
corticosteroid treatment and intensive insulin therapy for septic shock in adults
screening tests and all-cause mortality rapid MRI for patients with low back pain
follow-up of blood-pressure lowering and glucose control in type 2 diabetes . . .

In 2013 a clinical evidence group reviewed data on the effectiveness of 3,000 National Health Service treatments:

11% were rated beneficial
23% likely beneficial
7% trade-offs between benefits and harms
6% unlikely beneficial
3% likely ineffective or harmful
50% unknown effectiveness !

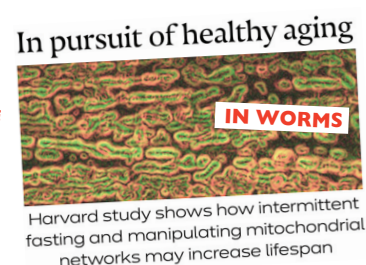
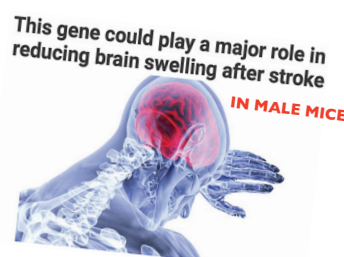
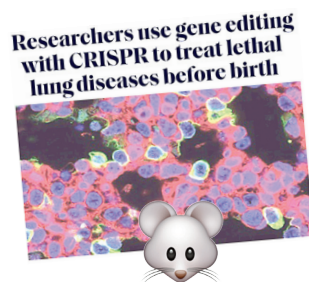
100%

Q.W. Smith, R.L. Street, R.J. Volk, M. Fordis, “Differing levels of clinical evidence,” *Medical Care Research Review*, 2013, 70, 3-13

⚠️ **CONSTRUCTING PREDICTIVE MODELS IS EASY, AUTHENTICITY AND PRACTICAL USE ARE DIFFICULT**

145 prediction models for covid-19 are “poorly reported, at high risk of bias, and their reported performance is probably optimistic. . . . The predictors identified could be considered as candidate predictors for new models. . . . Unreliable predictions could cause more harm than benefit in guiding clinical decisions.” *Similar problems occur in ~70% of all medical testing, procedures, research.*

Laure Wynants, et al, ‘Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal,’ *BMJ* 2020; updated April 5, 2020



⚠️ **“PROOF-OF-CONCEPT” AI MODELS ARE LIKE MOUSE STUDIES – EARLY RESULTS THAT CREATE FALSE HOPES FOR MEDICAL PATIENTS, MISLEADING PRESS RELEASES, PITCH SLIDES, PATENTS, COMMERCIAL EXPLOITATION. ALAS, RESEARCH AT THIS STAGE IS UNLIKELY TO EVER EXTEND LIVES.**

On the difference between machine learning and AI:

If it is written in Python, it's probably machine learning.

If it is written in Powerpoint, it's probably AI. MAT VELLOSO

A 2019 study identified 516 recently published articles using AI algorithms for medical image diagnosis. Did these studies use TRIPOD standard methods recommended for clinical validation of AI performance: (1) external/internal validation, (2) diagnostic cohort/case-control research designs, (3) data from multiple institutions, (4) prospective research designs – methods recommended for clinical validation of AI performance? The conclusion: “Only 6% of the 516 studies showed proof-of-concept technical feasibility, necessary for validation of AI for clinical work.” Also, some studies are compromised by leakage in high-dimensional spaces between model-building data and training data.

Dong Wook Kim, et al, ‘Design Characteristics of Studies Reporting Performance of Artificial Intelligence Algorithms for Diagnostic Analysis of Medical Images,’ *Korean Society of Radiology* 2019, edited.

BENEFITS OF VALIDATION, OPEN SOURCE, ETHICAL CHOICE OF PROBLEMS

Good research is rigorous and relevant, producing credible results that improve lives. Scientific credibility begins with honest independent quality assessments and replication with multiple data sets, especially for models with high-dimensional inputs. Half of all medical diagnostics and treatments are unvalidated, invalidated, or ineffective. These low-value practices are major targets of opportunity. Validated AI and machine learning are existential threats, and rightly so, to medical interventions with no better evidence than “we’ve always done it this way.” AI/machine learning should avoid arcane proprietary models, closed data, secret sauces; avoid harming or bankrupting patients; and avoid the EHR business model. Such avoidances would provide competitive advantages over most existing diagnostics. Beneficial research should (1) not falsify, (2) have longer time-horizons than press releases, (3) turn replicated research into diagnostic *and* treatment protocols, (4) reduce medical care costs, (5) use open source code and open-published research articles.

TRUTH AND CONSEQUENCES: META-RESEARCH AND EVIDENCE-BASED MEDICINE

	MARKED ENTHUSIASM	MODERATE ENTHUSIASM	NO ENTHUSIASM
RESULTS OF 6 WELL-DESIGNED (RANDOMIZED CONTROLLED) STUDIES:	0	3	3
RESULTS OF 47 POORLY DESIGNED STUDIES:	34	10	3

THOMAS CHALMERS, a founder of evidence-based medicine, demonstrated *the more susceptible a research design is to bias, the more enthusiastic the evidence for the favored treatments*. Replicated 1000s of times for medical interventions, this finding has survived for 50 years. For example, Chalmers and colleagues examined 53 published reports evaluating a surgery, portacaval shunts for esophageal bleeding. All reports were rated on (1) *enthusiasm* of findings favoring the surgery, (2) *quality of the research design* (good design = random assignment of patients to treatment or control; bad = treatment group not compared to any proper controls). The best design standard is the randomized controlled trial (RCT), which assigns patients randomly to the treatment or control group (assuring within chance limits that both groups are identical in all respects, known and unknown, and thereby avoiding, for example, selection of more promising patients to favored treatments). Of 53 published studies, only 6 *were well-designed (RCT)*, and **none** were markedly enthusiastic about the operation. In contrast, for 47 *reports lacking valid controls*, 72% enthusiastically endorsed a procedure unwarranted by the RCT standard. (ET, *Beautiful Evidence*, 145, revised)

JOHN IOANNIDIS published his classic paper, “Why Most Published Research Findings Are False” in 2005. Research-on-research has since flourished, quantifying what the word “most” means – which ranges from 40% to 98% depending on medical specialties. Observational studies appear atop the leaderboard of untrue results. Randomized controlled double-blind studies are by far most likely to be true, as Thomas Chalmers decisively proved. Improving the credibility and integrity of research might extend millions of statistical lives by favoring better treatments and by avoiding unnecessary, ineffective, harmful, costly treatments.

Why Most Published Research Findings Are False

John P. A. Ioannidis August 30, 2005

75,143
saves

5,814
citations

2,698,716
views

7,704
shares

(July, 2021)

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true

- when the studies conducted in a field are smaller
- when effect sizes are smaller
- when there is a greater number and lesser preselection of tested relationship
- where there is greater flexibility in designs, definitions, outcomes, analytical modes
- when there is greater financial and other interest and prejudice
- when more teams are involved in a scientific field in chase of statistical significance.

Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, claimed research findings may often be simply accurate measures of the prevailing bias. !

IF MOST PUBLISHED MEDICAL RESEARCH IS FALSE, THEN WHAT ABOUT META-RESEARCH?

Meta-researchers do relevant and sometimes stunning replication studies: for example, work by Glenn Begley and Lee Ellis reports that only 11% of 63 landmark preclinical studies of cancer drugs could be reproduced, and by Brian Nosek and colleagues evaluating the “Replicability of Social Science Experiments in *Nature* and *Science*, 2010 to 2015.”

Meta-research examines published journal articles – well-defined, stable, accessible populations. No missing data, no drop-outs, no randomization noise.

Meta-research is straight-forward, replicable, fast – (1) collect the research population, (2) identify bias, mess-ups, faults, falsification, validation issues, (3) replicate/confirm/adjudicate measurements (which, however, are often unblinded), (4) calculate prevalence of virtues and sins. Some measurements require judgments that might vary from judge to judge, such as ‘spin’ in abstracts, tactical citations (cites to other articles that are claimed to support the research findings, but in fact do not), whether subgroup analysis is worthy exploration or p-hacking. Precise definitions, independent replications, and weighted scoring may resolve such issues. Some measurements, however, are exact and replicable: specious accuracy, inappropriate image duplication, financial conflicts, binning, use of biomarkers not real outcomes, relative risks, and low-quality low-credibility research designs used in diet/nutrition studies, prediction modeling, oncology drug trials, and research on telepathy and flying saucers.

Meta-researchers assess credibility better than journal editors and their referees/reviewers. Speciality “reviews” are done by insider medical specialists, and usually fail to consider competing interventions beyond speciality fields, recognize new ideas, and rarely consider costs. Low-value research journals have published 1000s of sponsored, financially conflicted, biased, pseudo meta-research papers. Meta-researchers have a list of 800 ways to go wrong, sometimes with estimates of how much is wrong, how often it happens, when it matters. But empirical issues are turned into assumptions as in the slogan “unreliable evidence could cause more harm than benefits.” How much harm? To whom? Under what conditions? What are the economic costs and opportunity costs of untrue studies? Thus meta-analysis teams should include health economists.

Conflicts of Interest A multiplicity of motives drives research: sincere beliefs that their idea will benefit the world – and also advance their careers via publications, consulting gigs, making big money for their sponsors and themselves. Meta-research has few financial conflicts because there is no money there.

Meta-research is often descriptive – identify a problem, assess its prevalence. But how can lousy/misleading research be quarantined, stopped, quickly identified? PubPeer and select Twitter groups have recently proved effective. Retractions take many years and lawyers.

The Grand Truths of Meta-Research

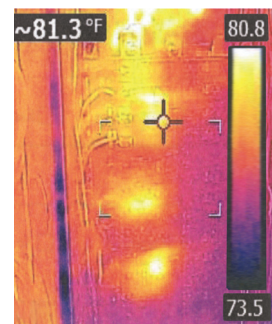
Well-designed randomized controlled trials are diamonds, many observational studies are quick sand. Confirmation bias is omnipresent. Money doesn’t talk, it screams. It’s more complicated than that.

Based on an inspection this day, the items marked below identify the violations in operation or facilities which must be corrected by the date specified below.

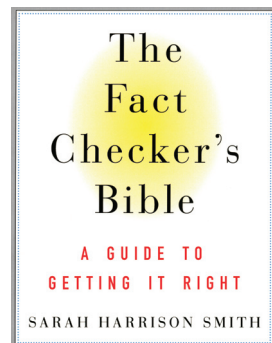
33

SOURCES OF FOOD			EQUIPMENT & UTENSILS : CLEANLINESS			VERMIN CONTROL		
1	Approved source, wholesome, nonadulterated	4	21	Preflushed, scraped, soaked and racked		42	Presence of insects/rodents	2
2	Original container, properly labeled	1	22	Wash water clean, proper temperature	1	43	Outer openings protected against entrance of insects/rodents	1
FOOD F 3 Pot: temp prep trans 4 Adeq temp			UNTRUE MEDICAL RESEARCH IS A MAJOR PUBLIC HEALTH PROBLEM: CREDIBILITY ASSESSMENTS OF RESEARCH SHOULD BE AS RIGOROUS AND FREQUENT AS RESTAURANT SANITATION INSPECTIONS AND PRE-SALE HOME INSPECTIONS			1		
5	Potentially hazardous food properly thawed	2	28	Equipment/utensils, storage, handling	1	46	Exterior walking, driving surfaces, good repair, clean	1
6	Unwrapped or potentially hazardous food not reserved	4	WATER SUPPLY			49	Walls, ceilings attached, equipment properly constructed, good repair, clean. Wall & ceiling surfaces as required.	1
7	Food protected during storage, preparation, display, service & transportation	2	29	Water source adequate, safe	4	50	Dustless cleaning methods used, cleaning equipment properly stored	1
8	Food containers stored off floor		30	Hot and cold water under pressure, provided as required	2	LIGHTING & VENTILATION		
9	Handling of food minimized	2	SEWAGE DISPOSAL			51	Adequate lighting provided as required	1
10	Food dispensing utensils properly stored	1	31	Sewage disposal approved	4	52	Room free of steam, smoke odors	1
11	Toxic items properly stored, labeled, used	4	32	Proper disposal of waste water	1	53	Room & equipment hoods, ducts, vented as required	
PERSONNEL			PLUMBING			DRESSING ROOMS & LOCKERS		
12	Personnel with infection restricted	4	33	Location, installation, maintenance	1	54	Rooms adequate, clean, adequate lockers provided, facilities clean	1
CLEANLINESS OF PERSONNEL			34	No cross connection, back siphonage, backflow	4	HOUSEKEEPING		
13	Handwashing facilities provided, personnel hands washed, clean	4	TOILET FACILITIES			55	Establishment and premises free of litter, no insect/rodent harborage, no unnecessary articles	1
14	Clean outer clothes, effective hair restraints	1	35	Adequate, convenient, accessible, designed, installed	4	56	Complete separation from living/sleeping quarters and laundry	1
15	Good hygienic practices, smoking restricted	2	36	Toilet rooms enclosed with self-closing door	1	57	Clean/soiled linens stored properly	1
EQUIPMENT & UTENSILS: DESIGN, CONSTRUCTION & INSTALLATION			37	Proper fixtures provided, good repair, clean		58	No live birds, turtles, or other animals (except guide dogs)	1
16	Food-contact surfaces designed, constructed, maintained, installed, located	2	HANDWASHING FACILITIES			SMOKING PROHIBITED		
17	Nonfood-contact surfaces designed, constructed, maintained, installed, located	1	38	Suitable hand cleaner and sanitary towels or approved hand drying devices provided, tissue waste receptacles provided	1	59	Smoking prohibited, signs posted at each entrance	3
18	Single service articles, storage, dispensing	2	GARBAGE/RUBBISH STORAGE & DISPOSAL			QUALIFIED FOOD OPERATOR		
19	No reuse of single service article		39	Approved containers, adequate number, covered, rodent proof, clean	1	60	Qualified Food Operator	3
20	Dishwashing facilities approved design, adequately constructed, maintained, installed, located	2	40	Storage area/rooms, enclosures – properly constructed, clean		61	Designated alternate	2
			41	Garbage disposed of in an approved manner, at approved frequency	1	62	Written documentation of training program	2
DEMERIT SCORE			RISK FACTOR VIOLATIONS IN RED					
4	3	2	Signature of Person in charge					
TOTAL			SIGNED (Inspector)					
RATING								
Date Corrections Due								

Medical research reports, directly relevant medical care and clinical guidelines, must be *independently* inspected – as carefully and frequently as health/sanitation officials inspect restaurants, hairdressers, water supplies, sewage systems. Or, in selling a house, where both parties and a mortgage bank bring in independent inspectors. This 42-page pre-sale report shows 55 photos, including infrared images of electrical switch boxes. Or, how about fact-checking of medical research as done routinely in serious news reporting and some nonfiction publications?



Performance audits and evaluations are always in danger of being distorted by financial interests, by compromised and corrupt bureaucracies and politicians, and by all of us who don't like arbitrary, fussy, inefficient, slow-motion bureaucracies telling us what to do. *But regulatory failure can turn into a massacre (270,000 U.S. deaths are attributed to oxycodone).* Every single oxycodone pill was approved by the U.S. Food and Drug Administration, and was made by licensed drug companies, prescribed by licensed doctors, sold by licensed pharmacists. All 72,000,000,000 pills (500 pills/U.S. household) were tracked to the exact place/time/amount of sale by the Drug Enforcement Agency. The only thing worse than regulatory agencies is the *Regulatory Theater of failed regulatory agencies captured, owned, and corrupted by those they regulate.*



QUICK CREDIBILITY SCORING OF RESEARCH ON HUMANS

substantive explanatory theory is vague, scientifically impoverished -7 randomized controlled trial +7
 failure to report relative *and* absolute risk in the same paragraph -5 empirical assessment of measurement error +4
 spurious correlation (eg, income drives both alleged cause and effect) -8 over-fitting -4
 unconflicted funding of research +5 independent honest validation/replication +8
 forensic audit of spreadsheet +5 undeclared financial conflict of interests -1 per \$25,000 to each author
 diet/nutrition study -5 contractor/sponsor/researcher have prior history of publication bias/retractions -7
 inappropriate image duplication -5 failed randomization -6 binning -5 severe multicollinearity -5
 summary models shown without underlying data -5
 p-hacking/model-hacking/subgroup-hacking -7 midcourse changes in research protocol -3
 ghostwritten or ghostgraphics -8 standard statistical model as sole assessment of error and uncertainty -6
 use of unusual, magical, arcane statistical methods (pseudo-control groups, poorly chosen
 instrumental variables, inappropriate cross-over designs, etc) -5 results dependent on model assumptions -6
 partied with sponsor's sales reps at professional meetings -0.5 each party, each author
 and other high-prevalence pitfalls and biases

Scoring elements are based on universal principles of scientific inference. Thus epidemiologists, meta-researchers, and unconflicted biostatisticians should design scoring elements based on severity and prevalence. Fair scoring is global, neutral, indifferent to specialties and disciplines (no local, private, unscientific definitions of causality, patient outcomes, financial conflicts). To avoid inevitable attempts to game/evade/commercialize scoring systems, and to avoid the capture of scoring by those scored ('we're all on the same team'), an independent *Consumer Reports* model might be the best defense.

SCORING RESEARCH DESIGNS AND DATABASES

The same data from the same research design is often published multiple times; this may indicate highly productive research or salami-slicing or vanity over-publishing. Recall the case of compromised randomization in the Mediterranean diet study, with 267 published follow-up studies that relied on the same database – score one, score all 267. Database and research design credibility scores carry over each time the data are published (discounting for local over-fitting and sub-sub-group analysis, etc). This reduces scoring costs.

More than half of credibility scoring can be automated. Financially unconflicted biostatisticians, meta-researchers, and epidemiologists can do the rest. Medical specialists have already had their say in the reviewing and editing of journal articles of their fellow specialists. Scoring should be replicated independently; if scores diverge, a third or even fourth scoring can adjudicate.

Research credibility and integrity scores can be totaled up for all sorts of interesting clusters: high-impact articles, research teams, specialties, sub-specialties, sponsors, journals, paywalled versus open source, publishers, university research centers, laboratories, guidelines – and the articles, databases, and research designs referenced/reported in new drug approvals.

**FROM RESEARCH TO DAILY PRACTICE: MEDICAL GUIDELINES HAVE SERIOUS CREDIBILITY ISSUES,
AND MAY CAUSE/JUSTIFY LUCRATIVE FALSE ALARMS, OVER-DIAGNOSIS, OVER-TREATMENT.**

People of the same trade seldom meet together, even for merriment and diversion, but the conversation ends in a conspiracy against the public, or in some contrivance to raise prices. ADAM SMITH

It is difficult to get people to understand something, when their salary depends on not understanding it. UPTON SINCLAIR

I've worked with 1000s of experts in psych diagnosis. Not one ever said 'Let's tighten criteria for my favorite diagnosis.' All worried about missed cases, none about harms and risks of mislabeling. ALAN CASSELS

It is simply no longer possible to believe much of the clinical research that is published, or to rely on the judgement of trusted physicians or authoritative medical guidelines. MARCIA ANGELL

Professional Societies Should Abstain From Authorship of Guidelines and Disease Definition Statements

John P.A. Ioannidis *Circulation*, October 2018 edited

“Guidelines from professional societies are increasingly influential. These documents shape how disease should be prevented and treated and what should come within the remit of medical care. Changes in definition of illness increases by millions the number of people who deserve specialist care: hypertension, diabetes mellitus, composite cardiovascular risk, depression, rheumatoid arthritis, or gastroesophageal reflux. Similarly, changes in prevention or treatment options may escalate overnight the required cost of care by billions of dollars. Should field specialists prepare such influential articles?

Professional society documents are written exclusively by insiders. Joining guideline panels is considered highly prestigious and allocation of writing positions is a unique means to advance an expert's visibility and career within the specific medical specialty. Tens of thousands of society members then cite these articles. Writing guidelines promotes the careers of specialists, building sustainable hierarchies of clan power, boosting the impact factors of specialty journals, elevating the visibility of the sponsoring organizations and conferences that massively promote society products to attendees. But do they improve medicine or do they promote biased, collective, organized ignorance?

Most published guidelines have red flags: sponsoring by a professional society with substantial industry funding, conflicts of interest for chairs and panel members, stacking, insufficient methodologist involvement, inadequate external review, and noninclusion of nonphysicians, patients, and community members. After the 2011 Institute of Medicine report, several societies changed the composition of their panels to avoid florid financial conflicts and preclude direct industry funding in guideline development. They have also included some methodologists. In recent guidelines, cardiovascular societies have tried to include more primary care

physicians, nurses, patients on their panels. However, it is unclear such representatives can exert much influence when embedded within a dominant majority of vocal specialists. Moreover, stacking of panels with specialists who have overt preferences is more difficult to avoid.

Some professional societies are huge financial enterprises. Producers of medical guidelines and disease definitions tend to be the largest financial players, with cardiology the leading example. For example, the annual American Heart Association budget in the fiscal year 2016-2017 was \$912 million, 20% of which came from corporate support. 77% of 60 million Euro annual income of the European Society of Cardiology comes from industry. Securing objectivity is difficult when industry-manufactured interventions also procure much of the specialty income. *Would a society therefore advise its members to change jobs, if evidence proved their medical services a waste?*

An overspecialized worldview is a major disadvantage in making sound recommendations. Specialists do not compare their merchandise against the merchandise of other healthcare providers. Specialists and societies compete for the same pie of healthcare resources.

Different countries vary on guidelines being entrusted to government or professional societies. In the United Kingdom, the National Institute for Health and Clinical Excellence is authorized by the government to consider both efficacy *and* cost. The US Preventive Services Task Force is convened by the Agency for Health Research and Quality, but most powerful guidelines are issued by professional societies; these place less attention on cost containment. With skyrocketing healthcare expenditures, largely cost-unconscious guidelines make little sense.”

“An alternative approach: specialists should not assume any major role in guidelines pertaining to their own fields. Instead of having mostly or exclusively specialists write the guidelines and occasional nonspecialists consult or comment, guidelines could be written by methodologists and patients, with content experts consulted and invited to comment. This approach has been proposed also for systematic reviews and meta-analyses that synthesize the evidence feeding into guideline development. Another possibility is to recruit to the writing team other medical specialists who are unrelated to the subject matter. Involvement of such outsiders (for example, family physicians) could be refreshing. These people may have strong clinical expertise, but no reason to be biased in favor of the specialized practices under discussion.

They may scrutinize comparatively what is proposed, with what supporting evidence, and at what cost. Devoid of personal stake, they can compare notes to determine if this makes sense versus what are typical trade-offs for evidence and decisions in their own, remote specialty. For example, while insider specialists may be willing to endorse an effective but highly expensive drug or device, outsiders may see more easily that this intervention is outrageously expensive. What may seem crucially important to a field expert, may appear as minutiae to a less personally involved outsider. Methodologists, patients, and different field specialists add to guideline teams more methodological rigor, patient-centeredness, and impartiality.”

SPECIALTY GUIDELINES GOVERN YOUR MEDICAL CARE

Internal medicine doctor: “Esophagram shows C7 spinal osteophyte. Nothing to worry about.”

Patient (ET): “Should a spine surgeon take a look?”

Internal medicine doctor: “No. He will operate on you.”

Patient thinking: [“That’s the best medical advice I have ever received.”]

When you choose a specialty, you choose diagnosis, treatment – and specialty guidelines. If you think that’s not a problem, read the definitive National Academy of Sciences Institute of Medicine report, *Clinical Practice Guidelines We Can Trust* or read the Ioannidis excerpt again.

RESISTANCE AND RESENTMENT TOWARD STATISTICAL EVIDENCE

“I do angioplasty and I have grateful patients. It’s not rocket science for me to figure out if PCI [percutaneous coronary intervention] works or not,” the chair of a leading cardiology department

Of course patients are grateful after an intervention – still alive, pain gone, high on anesthesia, told “you did great, everything went fine.” Post-op patient gratitude, applause ending PCI theater, does not prove an intervention actually works or is worth the immense cost compared to untheatrical alternatives (exercise, diet, drugs). Self-congratulatory measurements made by those keeping their own score are contrary to the proven truth, replicated 1000s of times: *poor research designs create false, but desired, findings*. Huge gains in cardiac health have come from *randomized controlled trials* that validated life-extending statins and from big data demonstrating the enormous cardiac health benefits of smoking cessation. Do specialists reject this evidence for being insufficiently anecdotal?

High levels of short-term patient satisfaction appear to be associated with hospitality (greeters at the door, empathetic staff, comfortable rooms) – but also with *more treatments, higher costs, substantially higher mortality even after adjusting for baseline health and comorbidities*.^{*} Several plausible stories might explain these observational findings. But immediate post-treatment patient satisfaction/gratitude does not measure whether the patient lives better and longer, the ultimate goal. Can placebo effects be produced more cheaply and less dangerously than a \$35,000 stent?

^{*}Note titles of these reports: Joshua J. Fenton, et al, “The Cost of satisfaction: A national study of patient satisfaction, health care utilization, expenditures, mortality,” JAMA Internal Medicine, 2012; and Cristobal Young and Xinxiang Chen, “Patients as consumers in the market for medicine: The halo effect of hospitality,” Social Forces, 2020

‘One day when I was a junior medical student, a very important Boston surgeon visited the school and delivered a great treatise on a large number of patients who had undergone successful operations for vascular reconstruction. At the end of the lecture, a young student at the back of the room timidly asked, “Do you have any controls?” Well, the great surgeon drew himself up to his full height, hit the desk, and said, “Do you mean did I not operate on half of the patients?” The hall grew very quiet then. The voice at the back of the room hesitantly replied, “Yes, that’s what I had in mind.” Then the visitor’s fist really came down as he thundered, “Of course not. That would have doomed half of them to their death.” God, it was quiet then, and one could scarcely hear the small voice ask, “Which half?”’ E. E. PEACOCK

SELF-EVALUATION, CONTEMPT FOR DATA, FINANCIAL CONFLICTS

Investigative reporting by ProPublica/*The New York Times* discovered that financial conflict of interest statements made by doctors and researchers in their published articles diverged from their *actual* conflicts – serving on drug company Boards of Directors (a fiduciary, primary loyalty), creating private start-ups, receiving huge consulting payments. Some deans and researchers soon departed their positions, and 1000s of corrections were made to previously published research papers. The Memorial Sloan Kettering physician-in-chief resigned 3 days after the report came out. Then, in a *New York Times* interview, the MSK replacement physician-in-chief defended financial conflicts by first-person experiences:

‘I’m telling you, as someone who works with patients, and I’ve worked with patients throughout my entire career here, that working with industry has helped me save lives. Maybe we should turn this around and say, we have more people on corporate boards because people value the opinions from our faculty.’

The Memorial Sloan Kettering chief of biostatistics/epidemiology had a different view:

‘Bias can creep into the scientific enterprise in all sorts of ways. But financial conflicts are detectable definitively and represent a uniquely perverse influence on the search for scientific truth. The key substantive issue is that the problems we face were not caused by failures to disclose conflicts. The problems were due to the conflicts themselves. Making billions is not our mission. MSK is a nonprofit with a fundamentally social mission.’

In response, the MSK replacement physician-in-chief sneered at biostatisticians and data:

‘He is a biostatistician. He lacks a full understanding of conflicts of interest. He does not work with patients. He works with data.’

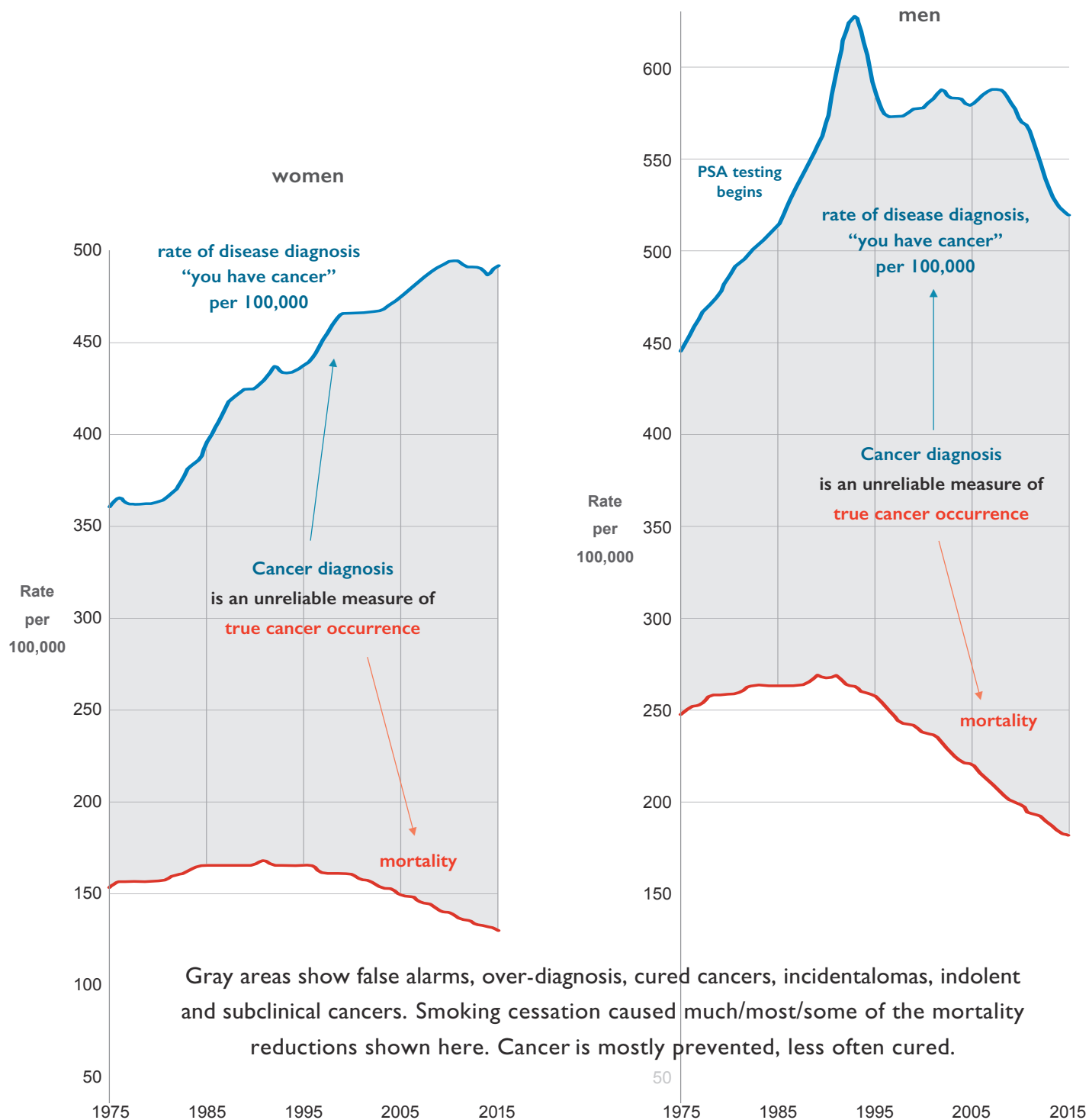
WORKING WITH DATA,

BIOSTATISTICIANS AND EPIDEMIOLOGISTS HAVE EXTENDED BILLIONS OF LIVES

Using vast amounts of data and intense detective casework in the field, epidemiologists measure new threats, then help respond, often successfully, to those threats. Their work, especially in vaccines and epidemics, has extended millions of lives and created billions of quality-life years. Randomized controlled trials, designed and analyzed by biostatisticians, identify interventions that extend lives. ‘The most important medical advance in our generation is not a pill, or a stent, or a surgery, but the randomized controlled trial,’ said Vinay Prasad. Statistical analysis proved that smoking causes cancer—leading to smoking cessation policies, preventing hundreds of millions of early deaths.

WALKING IN THE SHOES OF STATISTICAL LIVES:
40 YEARS OF CANCER DATA DESCRIBES MILLIONS OF INDIVIDUAL PATIENT EXPERIENCES,
EVALUATES CANCER SCREENING TESTS AND SUCCESS OF CLINICAL CARE

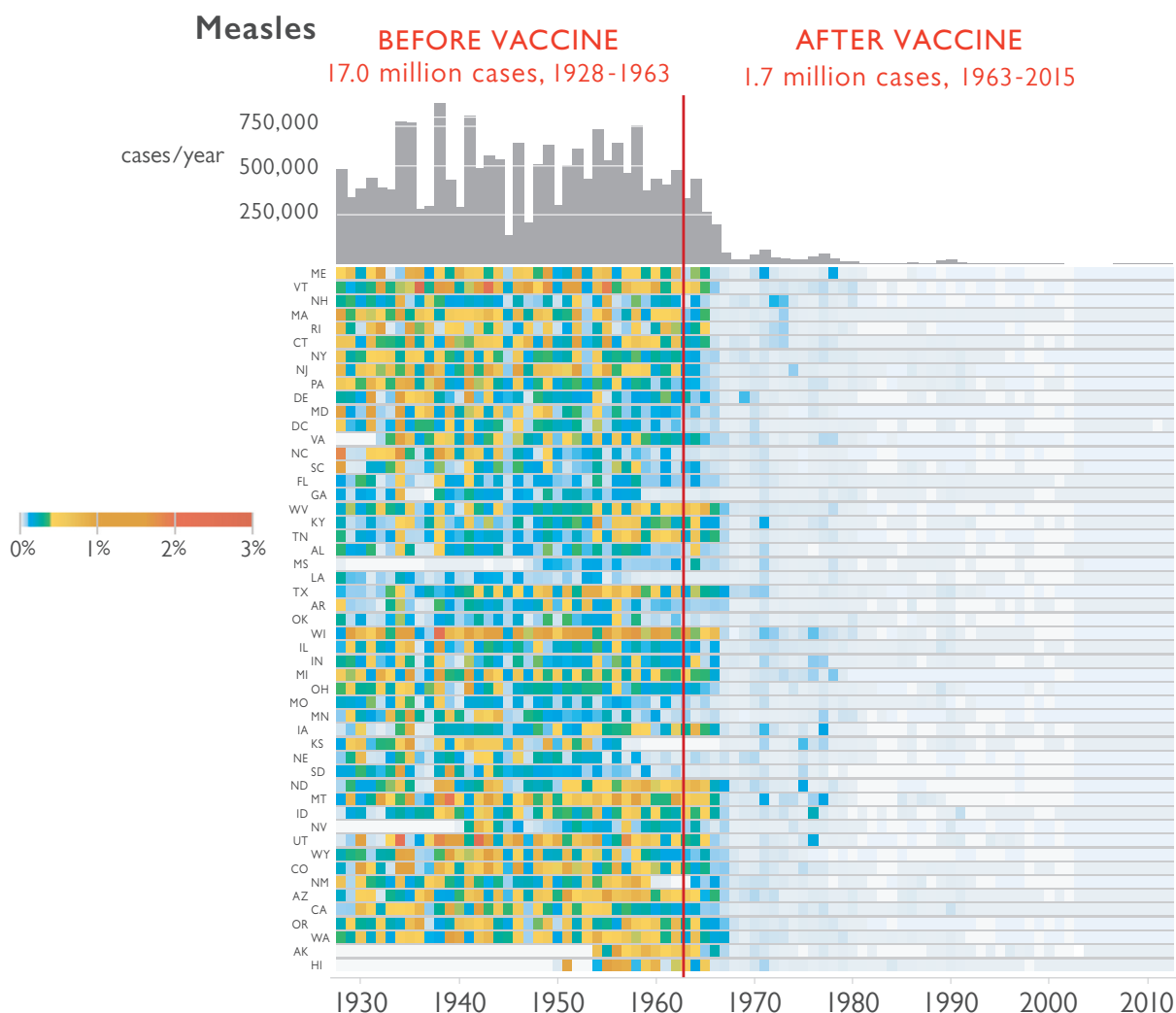
All cancers, rates of **disease diagnosis** (“you have cancer”) and **mortality**, U.S., 1975-2015



“Epidemiologic Signatures in Cancer,” Gilbert Welch, Barnett S. Kramer, William C. Black, *NEJM*, October 2019, graphics redrawn. See also *Malignant: How Bad Policy and Bad Evidence Harm People with Cancer*, Vinayak K. Prasad, 2020.

**KEEPING SCORE BY MEASURING STATISTICAL LIVES VALIDATES
INDIVIDUAL VACCINATIONS, PREVENTS 103 MILLION DISEASE CASES**

Building a data set of 88 million instances of disease each located in space and time from 1888 to 2011, epidemiologists were able to track the benefits of vaccination: a total of 103 million disease cases were prevented since 1924 in the United States. This natural experiment (before vs. after) shows strong consistent effects in all U.S. states for 5 diseases (measles, polio, rubella, hepatitis-A, mumps), as each state serves as its own control over the years – all adding up to direct visual proof.



Redrawn graphics based on Willem G. van Panhuis, et al., 'Contagious Diseases in the United States from 1888 to the present,' *New England Journal of Medicine* 369, 2013, 2152-2158; and Tynan DeBold and Dov Friedman, 'Battling Infectious Diseases in the 20th Century: The Impact of Vaccines,' *The Wall Street Journal*, 11 February 2015, based on original data from the *Morbidity and Mortality Weekly Report*, Centers for Disease Control, compiled and analyzed by Project Tycho at the University of Pittsburgh.

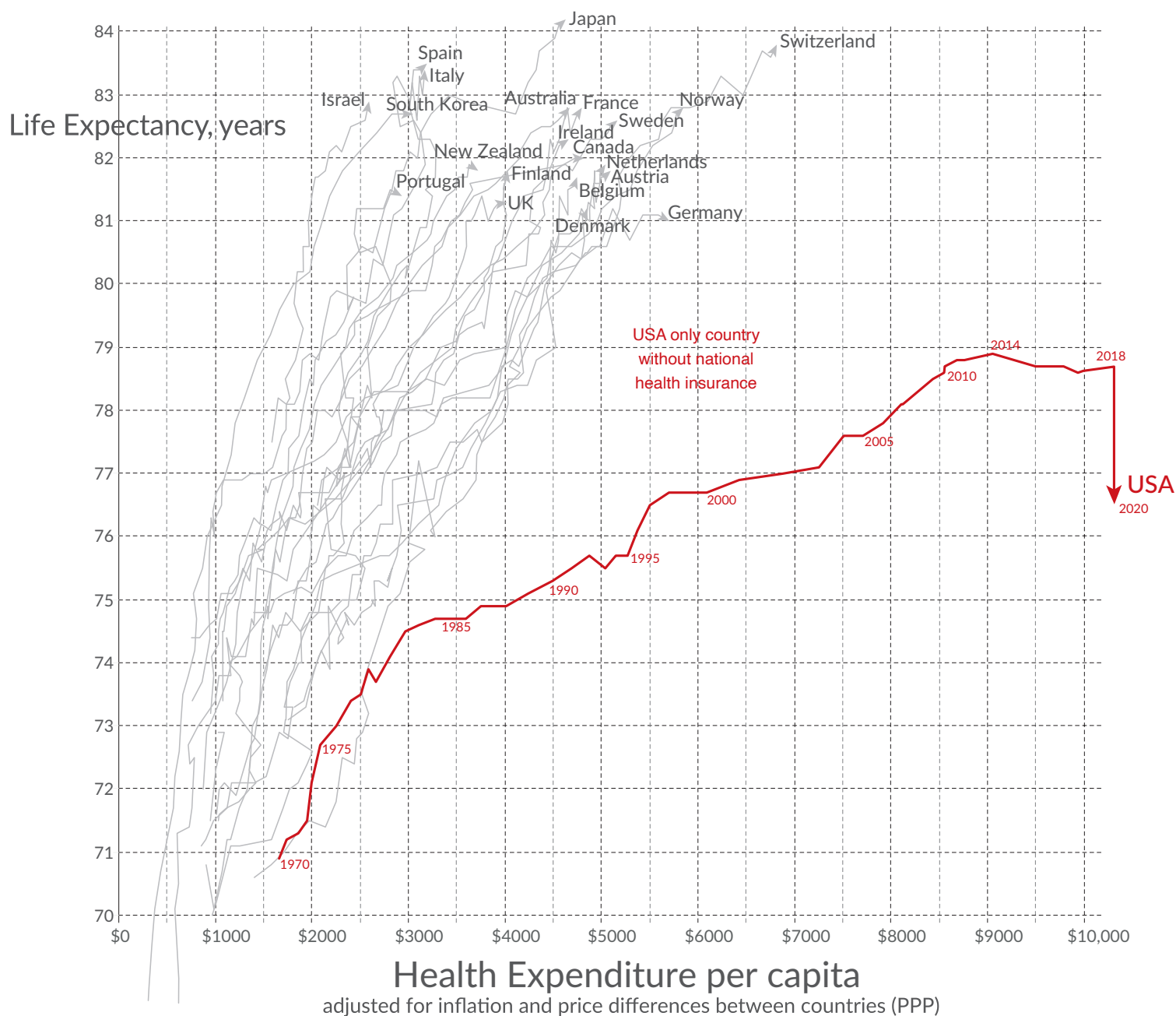
~4,000,000,000 STATISTICAL LIVES: DIVERGENT PERFORMANCE

TRAJECTORIES IN A 2-DIMENSIONAL SPACE, 33 COUNTRIES, 1970 TO 2020

Life expectancy vs. health expenditure

From 1970 to 2018

Our World
in Data



Data source: OECD — Note: Health spending measures the consumption of health care goods and services, including personal health care (curative care, rehabilitative care, long-term care, ancillary services, and medical goods) and collective services (prevention and public health services as well as health administration), but excluding spending on investments. Shown is total health expenditure (financed by public and private sources). Licensed under CC-BY by the author Max Roser.

OurWorldinData.org – Research and data to make progress against the world's largest problems.