HISTORY OF CONTROLLED EXPERIMENTS

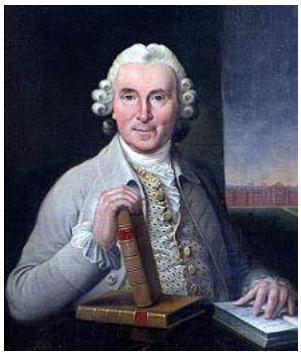
RON KOHAVI

UPDATED 15 NOV 2019

Commented [RK1]: See https://experimentguide.com/history/ for page with link to read-write version that you can comment on

I would love to get feedback on this brief history.
I tried to find the "first" of several key events.

The earliest controlled experiment appears to have been suggested in the Old Testament's Book of Daniel. King Nebuchadnezzar proposed that some Israelites eat "a daily amount of food and wine from the king's table." Daniel, an Israelite, preferred a vegetarian diet, but the official was concerned that the king would "see you looking worse than the other young men your age? The king would then have my head because of you." Daniel then proposed the following controlled experiment: "Test your servants for ten days. Give us nothing but vegetables to eat and water to drink. Then compare our appearance with that of the young men who eat the royal food, and treat your servants in accordance with what you see" (Daniel 1, 12– 13). This experiment fails many modern requirements, such as: sample size (treatment to four people), length (ten days is too short for a nutrition experiment), no control of the amount of food and exercise, the measurement is not well defined, there is no randomization, nor is there informed consent (Mosteller, Gilbert and McPeek 1983). Nonetheless, the key ideas are there!



James Lind
https://www.jameslindlibrary.org

Many centuries later, as long sea voyages became common and fresh fruits were rare on ships, an estimated two million sailors died of scurvy between 1500 and 1800 (Drymon 2008), a disease that today we know is caused by vitamin C deficiency. Dr. James Lind, a surgeon in the Royal Navy, observed the lack of scurvy among sailors serving on the naval ships of Mediterranean countries, where citrus fruit was part of their rations. In 1747, he conducted what is credited as being the first medical controlled experiment, testing six possible cures. On one voyage he gave some sailors oranges and lemons, and others alternative remedies including: vinegar, cider, and seawater. The sailors who ate oranges and lemons recovered quickly. While the experiment is often credited with being the first controlled experiment in medicine, it was a tiny experiment on a total of 12 sailors split into six treatments of two each. Dr. Lind did not appreciate his discovery and wrote about the experiment in only a few paragraphs in his book of over 500 pages, The Treatise of Scurvy (Lind 1753). It is likely that he was misled by his attempt to create a concentrated lemon juice by boiling juice from five pounds of oranges into five ounces of 'rob' extract. Unfortunately, vitamin C is destroyed under heat, and Dr. Lind continued to treat patients with other treatments, such as bloodletting. His experiment therefore had very little impact on addressing scurvy (Bartholomew 2002). Almost

50 years later in 1794, the ship Suffolk, which was fully supplied with lemon juice at the request of Rear Admiral Gardner, completed a four-month voyage without a trace of scurvy. The Admiralty supplied lemon to all ships from that point, with records showing that between 1795 and 1814, 1.6 million ounces of lemon juice were issued. British sailors are still called limeys today. Intelligent naval senior executive officers asserted that this event was the equivalent of doubling the fighting force of the navy (Bollet 2004). Lemons were in such high demand in the 1800s, that some have claimed that the Sicilian mafia appeared in locations where producers made high profits from citrus production (Dimico, Isopi and Olsson 2017). Interestingly, the interchangeable use of "lime" and "lemon" by the British led to tragedy, as the term lime juice was used for both the Mediterranean lemon and the West Indian lime; unfortunately, the latter is not an effective antiscorbutic. Some nineteenth and early twentieth expeditions to the Arctic and Antarctic carried the West Indian lime as their antiscorbutic, and thus suffered grievously from scurvy, whereas earlier expeditions, which had used lemons for this purpose, had been free of the disease (Bollet 2004).

Bloodletting was used by doctors for about two thousand years, from the first century BC to the mid-19th century, as a key therapy: open a vein in the arm with a knife called a lancet, and let blood out (Wooton 2007). The prevailing conception of illness was that the sick were contaminated by some toxin or contagion and that these conditions could be improved by opening a vein and letting the sickness run out—bloodletting. Once the toxins were gone, the patient immediately felt different, and often better. As anyone who has given blood knows, losing a pint or two makes you feel transformed. Intuitively, or correlationally, it was satisfying to doctors that the procedure left the patient quieter, screaming less. Bloodletting was used to treat almost every disease; a British medical text recommended bloodletting for acne, asthma, cancer, cholera, coma, convulsions, diabetes, epilepsy, gangrene, gout, herpes, indigestion, insanity, jaundice, leprosy, ophthalmia, plague, pneumonia, scurvy, smallpox, stroke, tetanus, tuberculosis, and for some one hundred other diseases (Carter and Carter 2005, p. 6). It was judged most effective to bleed patients while they were sitting upright or standing erect, and blood was often removed until the patient fainted.

Physicians often reported the simultaneous use of fifty or more leeches on a given patient. Through the 1830s the French imported about forty million leeches a year for medical purposes, and in the next decade, England imported six million leeches a year from France alone (Carter and Carter 2005, p. 7). President George Washington died after he had a throat infection, and three different doctors performed bloodletting and in total extracted more than half his blood volume in a short period. It is now believed that this procedure led to "preterminal anemia, hypovolemia, and hypotension (Vadakan 2004) and premature death of the first US president. In 1828, Pierre-Charles-Alexandre Louis published an article and then a book on the effects of bloodletting (Louis 1836, Morabia 2006). Louis took 77 patients from a very homogeneous group with the same, well-characterized form of pneumonia. He analyzed the duration of the

disease and the frequency of death by the timing of the first bloodletting (early in days 1-4, or later in days 5-9). The result: 44% of the patients who had been bled early died compared to 25% of those bled late. Bloodletting, he concluded was bad for you. He was so concerned about the result, that he opens the book as follows (Louis 1836):

> The results of my researches on the effects of bloodletting in inflammation, are so little in accordance with the general opinion, that it is not without a degree of hesitation I have decided to publish them. After having analyzed the facts, which relate to them, for the first time, I thought myself deceived, and began my work anew but having again from this new analysis, obtained the same results, I could no longer doubt their correctness.

Louis compared aggregated outcomes to treated groups, in what he called the "numerical method," which is recognized as an important precursor to clinical epidemiology and an early clinical trial, or a controlled experiment (Morabia 2006).

In 1835, a well-designed double-blind trial was done to evaluate the efficacy of a homeopathic dilution of salt and pure (snow) water vs a placebo of pure (snow) water alone (Stolberg 2006). Each of the 50 volunteers received a numbered vial, and the list of those vials with and without the salt dilution was sealed before the start of the trial. The volunteers had to report whether they perceived anything unusual. The homeopathy's proponent claimed the odds of experiencing extraordinary sensations were ten to one, yet only eight of the 50 participants reported anything unusual: five that received the dilution (Treatment) and three that received just water (Control). The homeopathic treatment was thus discredited. One interesting reason for challenging the study was that most participants have opposed homeopathy, and if they wanted to discredit it, they could simply report that they had not experienced anything unusual.

In 1908, William Sealy Gosset, a chemist working for the Guinness brewery in Dublin, Ireland, developed new statistical methods as Head Experimental Brewer. He was concerned with small samples and published the t-statistic, now heavily used in controlled experiments (Student 1908). Guinness did not allow their employees to publish, so Gosset published his work under the name "Student." Imagine the brand value for Guinness if instead of the "Student's t-test" it was called "Guinness' t-test."

https://en.wikipedia.org/wiki/Student%27s_t-test

https://en.wikipedia.org/wiki/Ronald_Fisher

The theory of a controlled experiment dates to Sir Ronald A. Fisher's experiments at the Rothamsted Agricultural Experimental Station in England in the 1920s. His book Statistical Methods for Research Workers (Fisher, Statistical Methods for Research Workers 1925) was one of the 20th century's most influential books on statistical methods where he introduced Fisher's method of meta-analysis and popularized p-values. This was followed by The Design of Experiments book (Fisher 1935), which argued for the use of randomized assignment. The "Lady Tasting Tea" is a famous randomized experiment devised by Fisher to evaluate whether a lady (Muriel Bristol) can correctly identify whether tea or milk was first added to a cup. Fisher proposed to give her eight cups in random order and see how many she could correctly identify (Fisher 1935, Salsburg 2002). Montgomery calls these years the "agricultural era" of four eras in modern statistical experimental design. (Montgomery 2012).

https://en.wikipedia.org/wiki/Austin_Bradford_Hill

In 1947-1948, the first published randomized and blinded clinical trial was done by at Great Britain's Medical Research Council (MRC) to assess the efficacy of streptomycin for treating pulmonary tuberculosis. 107 patients were assigned using random number assignments devised by Austin Bradford Hill, and the radiologists who interpreted the radiographs were also blind to the assignment (Doll 1998). While ethical questions were raised about whether to withhold a drug that had been effective in animal experiments, only a small amount of the drug streptomycin was available.
The Medical Research Council agreed that "[it would] have been unethical *not* to have seized the opportunity to design a strictly controlled trial which could speedily and effectively reveal the value of the treatment" (Hill 1963). In later years, R. A. Fisher's argument that randomization also allowed precise estimates of errors became accepted in Randomized Clinical Trials (RCTs).

It is interesting to contrast the successful execution by the MRC with the poor execution of a similar trial by the Veterans Administration, which did not follow protocols well (e.g., for random assignment). As Harry Marks (1988) noted from the Minutes of the First Streptomycin Conference in 1946: The very idea that they were conducting experiments had to be approached gingerly: "We don't like to use the word 'experiments' in the Veterans Administration;

'investigation' or 'observations,' I believe is the approved term for such a study in the VA hospitals."

The second era, called the "industrial era" (Montgomery 2012), was led by the development of response surface methodology (RSM) by Box and Wilson (1951). Unlike in agriculture, they focused on the fact that many industrial experiments have a nearly immediate observable OEC (response), and that the experimenter can learn and iterate with the next experiment (a feature called sequentiality). Both of these characteristics are commonly utilized in online controlled experiments.

https://en.wikipedia.org/wiki/George_E._P._Box

In 1970, the US FDA and the courts made controlled clinical trials, using a statistical model, a matter of regulatory law and legal precedent. Marks wrote that the "Adoption of controlled experimentation in medicine has generally been interpreted as a triumph for the intellectual power and cogency of statistical concepts and theory" (Marks 1988, 319).

The third experimentation era, called the "Robust parameter design" (Montgomery 2012) was led by quality improvements in industry, promoted by Genichi Taguchi with fractional factorial designs (Taguchi 1987).

A detailed history of randomized controlled trials (RCTs) to the early 1970s is summarized by Marcia Meldrum (2000). Harry Marks (1997) summarized RCTs from 1900 to 1990. One of the classical books for offline experiments is Statistics for Experimenters (Box, Hunter and Hunter 2005) whose first edition was in 1978.

We are now in the fourth era of self-service online experimentation platforms operating at scale.

ACKNOWLEDGMENT

Thanks to Shilpa Gupta for feedback on an earlier draft.

REFERENCES

Bartholomew, Michael. 2002. "James Lind and scurvy: A revaluation." *Journal for Maritime Research* 4:1: 1-14. doi:http://dx.doi.org/10.1080/21533369.2002.9668317.

Bollet, Alfred Jay. 2004. *Plagues & Poxes: The Impact of Human History on Epidemic Disease.* 2nd edition. Demos Health.

Box, G. E. P., and K. B. Wilson. 1951. "On the Experimental Attainment of Optimum Conditions." *Journal of the Royal Statistical Society. Series B (Methodological)* 13 (1): 1-45.

Box, George E.P., J Stuart Hunter, and William G Hunter. 2005. *Statistics for Experimenters: Design, Innovation, and Discovery.* 2nd. John Wiley & Sons, Inc.

Carter, Codell K, and Barbara R Carter. 2005. *Childbed Fever: A Scientific Biography of Ignaz Semmelweis.* Rutgers, New Jersey: Transaction Publishers.

Dimico, Arcangelo, Alessia Isopi, and Ola Olsson. 2017. "Origins of the Sicilian Mafia: The Market for Lemons." *The Journal of Economic History* 77 (4): 1083-1115. doi:10.1017/S002205071700078X.

Doll, Richard. 1998. "Controlled trials: the 1948 watershed." *BMJ.* doi:https://doi.org/10.1136/bmj.317.7167.1217.

Drymon, M. M. 2008. *Disguised as the Devil: How Lyme Disease Created Witches and Changed History.* Wythe Avenue Press.

Fisher, Ronald Aylmer. 1925. *Statistical Methods for Research Workers* .

—. 1935. *The Design of Experiments.*

Hill, Austin Bradford. 1963. "Medical Ethics and Controlled Trials." *British Medical Journal (BMJ)* 1043-1049. doi:https://doi.org/10.1136/bmj.317.7167.1217.

Lind, James. 1753. *A Treatise on the Scurvy.* London: A. Millar.

Louis, Pierre-Charles-Alexandre. 1836. *Researches on the Effects of Bloodletting in Some Inflammatory Diseases, and on the Influence of Tartarized Antimony and Vesication in Pneumonitis.* Forgotten Books reprint 2018. Translated by C. G. Putnam. Boston, MA: Hilliard, Gray, and Company.

Marks, Harry M. 1988. "Notes from the Undergound: The Social Organization of Therapeutic Research." In *Grand Rounds: One Hundred Years of Internal Medicine*, by Russell C Maulitz and Diana E Long, 297-336. University of Pennsylvania Press.

—. 1997. *The Progress of Experiment: Science and Therapeutic Reform in the United States, 1900-1990.* Cambridge University Press.

Meldrum, Marcia L. 2000. "A Brief History of the Randomized Controlled Trial: From Oranges and Lemons to the Gold Standard." *Hematology/Oncology Clinics of North America* 745-760. doi:https://doi.org/10.1016/S0889-8588(05)70309-9.

Montgomery, Douglas C. 2012. *Design and Analysis of Experiments.* 8 edition. Wiley.

Morabia, Alfredo. 2006. "Pierre-Charles-Alexandre Louis and the evaluation of bloodletting." *Journal of the Royal Society of Medicine* 99: 158-160.

Mosteller, Frederick, John P Gilbert, and Bucknam McPeek. 1983. "Controversies in Design and Analysis of Clinical Trials." In *Clinical Trials*, by Stanley H Shapiro and Thomas A Louis. New York, NY: Marcel Dekker, Inc.

Salsburg, David. 2002. *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century.* Holt Paperbacks.

Stolberg, M. 2006. "Inventing the randomized double-blind trial: the Nuremberg salt test of 1835." *Journal of the Royal Socienty of Medicine* 99: 642-643.

Student. 1908. "The Probable Error of the Mean." *Biometrika* 1-25.

Taguchi, Genichi. 1987. *The System of Experimental Design: Engineering Methods to Optimize Quality and Minimize Costs.* White Plains, N.Y.: UNIPUB.

Vadakan, Vibul V. 2004. "The Asphyxiating and Exsanguinating Death of President George Washington." *The Permanente Journal* 8 (2).

Wooton, David. 2007. *Bad Medicine: Doctors Doing Harm Since Hippocrates.* Oxford University Press.