BJHS Themes (2026), 1–17
doi:10.1017/bjt.2026.10034

RESEARCH ARTICLE

The history of statistics in statistically valid regulation

Christopher J. Phillips

History Department, Carnegie Mellon University, USA
Email: cphillip@andrew.cmu.edu

Abstract

This paper uses the introduction at the US Food and Drug Administration of dose–response extrapolation for ascertaining toxicity between 1950 and 1980 to analyse the negotiations between statistical knowledge and regulation. Those statistical methods enabled experimental results to be translated into a 'safe dose' of a substance for human consumption, but different methods continued to give different estimates of effects, and there was little basis for determining which methods were most accurate. I argue that statisticians were not proposing their discipline so much as a tool for mechanical decision making than as a set of methods for establishing a regulatory procedure that made assumptions and judgements visible. Consequently, their use did not bring debates about low-dose toxicity to a close even as they promoted a regulatory ethos and enabled regulators to act, even in cases of inherent uncertainty and inescapable variability.

Decision making concerning the regulatory approval of new pesticides or pharmaceuticals was rarely straightforward in twentieth-century America. Drugs were effective for some patients and not others; substances cause cancer in animals but not in humans; nearly any substance was toxic at high enough doses but still possibly safe and effective at the 'right' dose. Between 1950 and 1980, the US Food and Drug Association (FDA) turned to statistical methods developed largely by statisticians at the National Institutes of Health (NIH) as a solution to these problems of variability and uncertainty.

Though randomized clinical trials are perhaps the best-known and most-examined example of statistical methods entering mid-century health regulations, in fact there were far more cases in which statistical analysis of observational and experimental data were drawn on by regulators. This paper focuses on the relatively little-known case of statistical methods for 'low-dose' extrapolation. Even if a pesticide was carcinogenic for humans or animals at high doses, the effects at the low doses at which it was intended to be used were often much more difficult to measure. Animal experiments on potentially toxic substances, particularly ones which might be carcinogenic, were nearly impossible to interpret without statistical extrapolation methods. The presence of tumours in experimental subjects could be a result of chance; the absence of tumours was not a guarantee of safety. Moreover, doses at which toxicity could be reliably observed often needed to be translated into a measure of risk for the doses to which people would be likely to be exposed. Those statistical methods enabled experimental results to be translated into a 'safe dose' of a substance for human consumption, which in turn could be used, for example, to reverse-engineer the appropriate dosage of pesticides on crops.

© The Author(s), 2026. Published by Cambridge University Press on behalf of British Society for the History of Science. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives licence (http://creativecommons.org/licenses/by-nc-nd/4.0), which permits non-commercial re-use, distribution, and reproduction in any medium, provided that no alterations are made and the original article is properly cited. The written permission of Cambridge University Press or the rights holder(s) must be obtained prior to any commercial use and/or adaptation of the article.

https://doi.org/10.1017/bjt.2026.10034 Published online by Cambridge University Press

From the perspective of the history of medicine, the introduction of statistical methods into the clinic served as a crucial part of the machinery of reducing bias.[1] In the history of regulatory science, statistical methods make an appearance as tools deployed in rule making but not usually as a contested domain of scientific research in their own right.[2] Early STS scholars of science and regulation, perhaps most prominently Sheila Jasanoff, put statistical and probabilistic calculations of risk at the centre of many of the debates over regulatory science in the 1970s and 1980s. Jasanoff noted convincingly how, for example, the FDA successfully deployed statistical methods as part of a 'pragmatic approach' involving 'ad hoc procedures and arguments'.[3] Scholars like Jasanoff were less focused, however, on why those debates ultimately centred on disagreements about whether and how statistical methods should be used to make inferences in cases of variability and uncertainty. Those disagreements were not simply between statisticians and other scientists, but among statisticians themselves. As I will argue in this article, one clear area of agreement among statisticians was that the statistical methods of dose–response extrapolation did not ensure that any form of truth or accuracy would result from the calculations.

Below, I examine the case of dose–response modelling within regulatory science, showing how statisticians were consistently wary of using statistical methods to ascertain the precise effects of exposures to potentially carcinogenic substances. They focused on the use of statistics as a tool for regulation and validation not because statistical methods reliably provided accurate results but rather because statistical methods enabled assumptions and judgements to be made explicit and a consensus to be generated from aggregated evidence. Statisticians did not want to be sidelined as mere technicians or accountants; they wanted to have a seat at the table, collaborating on the often subtle interpretation of inescapably uncertain and variable medical data. Towards the end of the article, I briefly note the parallels with the better-known case of randomized clinical trials to gesture at the ways my arguments about the statistics of dose–response extrapolation might generalize to the use of statistics in regulation more broadly.

This paper is centred on US regulatory agencies partially because this is a regulatory setting in which analytical arguments are 'politically exposed' compared to other systems: regulators must justify decisions publicly and so the methods by which those decisions are made are visible and consequential.[4] But it was also in the American context that statistical regulation of experimental practices in medicine and health took initial form. It was increasingly unacceptable by the 1960s to subject humans to higher doses of drugs simply to determine at what level they became toxic, or to randomize groups to be exposed to a putatively toxic chemical simply to measure the response. Statistical extrapolation arose precisely in the period David Rothman once identified as critical for the circumscription of the 'authority' of physicians to do as they liked, from 1966 to 1976.[5] The NIH and US Public Health Service played central roles, both in the formulation of early guidelines for

[1] E.g. Harry M. Marks, *The Progress of Experiment: Science and Therapeutic Reform in the United States, 1900–1990*, Cambridge: Cambridge University Press, 1997.

[2] Soraya Boudia, 'From threshold to risk: exposure to low doses of radiation and its effects on toxicants regulation', in Soraya Boudia and Nathalie Jas (eds.), *Toxicants, Health and Regulation since 1945*, London: Routledge, 2013, pp. 71–87; Daniel Carpenter, *Reputation and Power: Organizational Image and Pharmaceutical Regulation at the FDA*, Princeton, NJ: Princeton University Press, 2010.

[3] Sheila Jasanoff, 'Science and the limits of administrative rule-making: lessons from the OSHA cancer policy', *Osgoode Hall Law Journal* (1982) 20, pp. 536–61, 541; see also Jasanoff, *The Fifth Branch: Science Advisers as Policymakers*, Cambridge, MA: Harvard University Press, 1990; Jasanoff, 'Science, politics, and the renegotiation of expertise at EPA', *Osiris* (1992) 7, pp. 194–217.

[4] Ronald Brickman, Sheila Jasanoff and Thomas Ilgen, *Controlling Chemicals: The Politics of Regulation in Europe and the United States*, Ithaca: Cornell University Press, 1985, pp. 303–10.

[5] David J. Rothman, *Strangers at the Bedside: A History of How Law and Bioethics Transformed Medical Decision Making*, New York: Basic Books, 1991, esp. p. 107.

https://doi.org/10.1017/bjt.2026.10034 Published online by Cambridge University Press

treatment in the NIH's Clinical Center and in the broader formulation of an ethical code of research.[6] The NIH, as the world's most substantial and influential funder and convener of medical research, played an outsized role in the global transformation of mid-century medicine; its statisticians likewise played a central role in the establishment of statistics for American regulations about health and medicine.

The statistics underlying dose–response models were not new in the postwar era. The practices of fitting data points to curves went back at least to the early nineteenth century, and the interpretation of those curves as doses and responses had been established in the 1930s. Chester Bliss, an entomologist who trained alongside statistician R.A. Fisher, studied the responses of insects to different doses of pesticides. Lethality varies among individuals, and he noted that it was difficult to predict the precise dose that might kill any given animal.[7] Bliss, instead of treating a particular dose level as binary (lethal/non-lethal), chose to think in terms of probability: what percentage of subjects is a particular dose likely to kill? This turned the question into one involving a cumulative probability distribution, enabling him to plot dose against percentage killed, and then predict the mortality rate even at doses not administered. His efforts, published in the mid-1930s, made it clear how to turn an experiment using different doses into a problem amenable to the 'maximum likelihood' methods that Fisher and others were promoting as the key to inferential statistical methods. The remaining debates were mainly about how to weigh evidence at different doses, particularly when there were different numbers or kinds of animals tested at different doses.[8]

If the methods were not new, what *was* new at mid-century was the explicit use of statistical methods in regulatory settings, an introduction overseen by a remarkably small group of statisticians who worked for the US federal government. For much of the nineteenth century and the early twentieth, statistical operations at the governmental level were largely limited to census tabulations as well as economic and vital statistics data.[9] Starting in the late 1940s, however, the Public Health Service created a unit of statisticians to aid with the experimental investigations that fell under the broad heading of biometry. Led by demographer and sociologist Harold Dorn within the 'methods' division of the Public Health Service, the unit would soon move to the National Cancer Institute and then be gradually dispersed across the National Institutes of Health as the agency expanded in the 1950s.[10] This relatively small group, including Jerome Cornfield, Nathan Mantel, Marvin Schneiderman, and Samuel Greenhouse, came from the world of demography, economics, and labour statistics, rather than from the biomedical fields, and they argued that the problems that regulators faced were ones of making inferences on the basis of limited information, rather than the traditional descriptive statistics of epidemiology and public health.[11] The problems of

[6] John A. Lynch, *The Origins of Bioethics: Remembering When Medicine Went Wrong*, East Lansing: Michigan State University Press, 2019, pp. 32–3, 36–7.

[7] C.I. Bliss, 'The calculation of the dosage–mortality curve', *Annals of Applied Biology* (1935) 22(1), pp. 134–67.

[8] J.O. Irwin and E.A. Cheeseman, 'On the maximum-likelihood method of determining dosage–response curves and approximations to the median-effective dose, in cases of a quantal response', *Supplement to the Journal of the Royal Statistical Society* (1939) 6(2), pp. 174–85; see also the discussion in J.O. Irwin, 'Statistical method applied to biological assays', *Supplement to the Journal of the Royal Statistical Society* (1937) 4(1), pp. 49–60.

[9] Joseph W. Duncan and William C. Shelton, *Revolution in United States Government Statistics, 1926–1976*, Washington, DC: US Government Printing Office, 1978; Duncan and Shelton, 'U.S. government contributions to probability sampling and statistical analysis', *Statistical Science* (1992) 7(3), pp. 320–38.

[10] Sejal Patel, 'The benevolent tyranny of biostatistics: public administration and the promotion of biostatistics at the National Institutes of Health, 1946–1970', *Bulletin of the History of Medicine* (2013) 87(4), pp. 622–47; Jonas H. Ellenberg, Mitchell H. Gail and Nancy L. Geller, 'Conversations with NIH statisticians: interviews with the pioneers of biostatistics at the United States National Institutes of Health', *Statistical Science* (1997) 12(2), pp. 77–81.

[11] Christopher J. Phillips, 'Modeling the epidemiologic individual', *History of the Human Sciences*, 6 June 2025, online first, at https://doi.org/10.1177/09526951251337680 (accessed 20 February 2026).

https://doi.org/10.1017/bjt.2026.10034 Published online by Cambridge University Press

surveys, sampling and demographic modelling involved statistical methods that were well suited to solving problems in dose–response models, clinical trials and observational studies. The period from roughly 1968 to 1973 was the pivotal one at the FDA, when the basic methods for clinical trials for new drug applications, as well as the methods of determining 'safe' doses of pesticides, were formalized. During those years, the head of Biometry and Epidemiology within the FDA's Bureau of Drugs was Charles Anello, himself a former student of Cornfield's at Johns Hopkins. Anello was a crucial instigator of the rapid expansion of the place of statistics at the FDA; he singlehandedly led the agency to expand from fewer than ten full-time statisticians at the Bureau of Drugs in 1970 to more than double that number by the end of the decade.[12] He was also the person primarily responsible for bringing in NIH statisticians like Cornfield to advise the FDA on statistical matters.

The statisticians developing methods for regulators remain even more invisible than the regulators themselves. Nevertheless, the introduction of statistical methods in the 1960s was pivotal, laying groundwork for the expansion of risk modelling and cost–benefit analysis across the 1970s and 1980s. Paying attention to the statistics and statisticians themselves helps provide a much more nuanced and rich understanding of the history of medicine and biostatistics at mid-century. In particular, government statisticians' approach to what would come to be known as quantitative risk analysis was very different from that of economists, and much more focused on a modest use of statistics for regulators.

Dose–response statistics: assigning risk to unobservable doses

When the initial group of statisticians arrived at the NIH campus in 1947, they were first focused on the question of dose response for the chemotherapeutic agents being tested at the National Cancer Institute.[13] Given the novelty of chemotherapeutic regimes, balancing harm and benefit was a matter of life and death. For the statisticians, however, the most pressing practical problem with dose–response models was the complexity of computation required, so the group took on that task first.[14] The problem that would prove more intractable was that of extreme doses. This problem had plagued statisticians right from the start, with Bliss's admission in the 1930s that changes in mortality at very low doses seemed qualitatively different from those at mid-range doses, though it wasn't at all obvious to Bliss whether this was a 'mathematical artifact' or 'biological reality'.[15] It was clear in practice that a finite dose might affect all the animals (and that there was some non-zero dose that would fail to affect any of them), even though the continuous probabilistic model implied that 100 per cent or 0 per cent effectiveness only occurred at an infinite or zero dose. Bliss had enlisted Fisher to provide a solution in an appendix to one of his original papers; Fisher chose to assign an arbitrarily small but non-zero response rate to doses for which no animal responded, thus enabling the dose–response curve to be calculated as usual.[16]

[12] Robert T. O'Neill, 'Reflections on professor Jerome Cornfield's contributions to the US Food and Drug Administration', *Clinical Trials* (2013) 10(2), pp. 332–6, 332; Charles Anello and Suzanne Junod, 'Contributions of Robert T. O'Neill to the evolution of regulatory statistical science', *Statistics in Biopharmaceutical Research* (2013) 5(3), pp. 262–79.

[13] Nathan Mantel, 'Jerome Cornfield and statistical applications to laboratory research: a personal reminiscence', *Biometrics* (1982) 38, pp. 17–23.

[14] Jerome Cornfield and Nathan Mantel, 'Some new aspects of the application of maximum likelihood to the calculation of the dosage response curve', *Journal of the American Statistical Association* (1950) 45(250), pp. 181–210, 181.

[15] Bliss, op. cit. (7), p. 144.

[16] Bliss, op. cit. (7), pp. 164–5.

https://doi.org/10.1017/bjt.2026.10034 Published online by Cambridge University Press

This statistical assumption was justified in part on the biological assumption that there was a threshold below which toxic doses would have no effect, but in the 1950s that biological assumption about thresholds was appearing increasingly untenable as even low levels of exposure seemed to pose risks in certain cases. After the atomic bombs were dropped on Hiroshima and Nagasaki in 1945, it was soon obvious that there were both short-term acute toxic effects for those closest to the detonation and long-term effects even for those exposed to much lower levels of radiation at a distance. Indeed, documentation of the health problems of the unlucky 1954 *Lucky Dragon* fishermen working near the testing site of a hydrogen bomb, as well as the 1957 identification of the radioactive element strontium-90 in children's teeth, raised alarms about exposure to tiny levels of radiation. In the early 1960s, Rachel Carson's *Silent Spring* highlighted the persistent and invidious effects of substances like DDT. Other novel chemicals, like thalidomide, had disastrous unintended consequences at reduced inadvertent exposures. Compounding the issue, some substances, like radiation and DDT, persisted in the environment, meaning that humans might be exposed to them for many years at small or hard-to-detect levels.[17]

The potential problem of low-dose exposures to hazardous substances was particularly acute with cancer. The newly dominant model of cancer at mid-century, 'somatic mutation theory', held that there was no lower threshold for exposure to carcinogens, with a risk for metastasis from cancer in even one cell. The theory, initially proposed in the 1940s, posited that cancer was a product of genetic changes (mutagenesis) in the ordinary cells and tissues of the body, turning normal cells into cancerous, or cancer-generating, cells. Researchers concluded that cancer was caused by changes in the cells that were cumulative and irreversible, and which spread through cell division, suggesting that the operative mechanism was mutagenesis.[18] Somatic mutation theory posited that there was no lower threshold for a mutation and the growing evidence in the 1950s from radiation research supported jettisoning the 'threshold' concept in favour of a model of 'progressive contamination'.[19] The International Commission on Radiation Protection concluded shortly thereafter that the best evidence supported a linear relationship across doses, with no lower or upper threshold, a cumulative effect of doses regardless of rate or interval, and irreversible effects. Taken together, there would be no safe level, which led the commission to focus on 'acceptable' or

[17] Among the large literature tracing mid-century concerns about toxicity see esp. Soraya Boudia, Angela N.H. Creager, Scott Frickel, Emmanuel Henry, Nathalie Jas, Carston Reinhardt and Jody A. Roberts, *Residues: Thinking through Chemical Environments*, New Brunswick, NJ: Rutgers University Press, 2021; Angela N.H. Creager and Jean-Paul Gaudillière (eds.), *Risk on the Table: Food Production, Health, and the Environment*, New York: Berghahn Books, 2021; Soraya Boudia and Nathalie Jas (eds.), *Powerless Science? Science and Politics in a Toxic World*, New York: Berghahn Books, 2014; Kate Brown, *Plutopia: Nuclear Families, Atomic Cities, and the Great Soviet and American Plutonium Disasters*, Oxford: Oxford University Press, 2013; Nancy Langston, *Toxic Bodies: Hormone Disruptors and the Legacy of DES*, New Haven, CT: Yale University Press, 2010; Sarah Vogel, 'From "the dose makes the poison" to "the timing makes the poison": conceptualizing risk in the synthetic age', *Environmental History* (2008) 13(4), pp. 667–73; John Wargo, *Our Children's Toxic Legacy: How Science and Law Fail to Protect Us from Pesticides*, 2nd edn, New Haven, CT: Yale University Press, 1998.

[18] Angela N.H. Creager, '"EAT. DIE." The domestication of carcinogens in the 1980s', in Creager and Gaudillière, op. cit. (17), 107; Heiko Stoff, 'Poison and cancer: the politics of food carcinogens in 1950s West Germany', in Creager and Gaudillière, op. cit. (17), Angela N.H. Creager, 'Radiation, cancer, and mutation in the atomic age', *Historical Studies in the Natural Sciences* (2015) 45(1), pp. 14–48; Creager, *Making Mutations Matter: A History of Environmental Risk through a Bacterial Test for Cancer-Causing Chemicals* (forthcoming); Volker Wunderlich, 'Zur entstehungsgeschichte der Druckrey–Küpfmüller-schriften (1948–1949): Dosis und Wirkung bei krebserzeugenden stoffen' (On the origins of the Druckrey–Küpfmüller papers (1948-1949): dose-effect relations in carcinogenic substances), *Medizinhistorisches Journal* (2005) 40(3–4), pp. 369–97; Wunderlich, '"Mit papier, Bleistift und Rechenschieber": Der krebsforscher Hermann Druckrey im internierungslager Hammelburg (1946–1947)' ('With paper, pencil, and slide rule': cancer specialist Hermann Druckrey in internment camp Hammelburg (1946–1947)), *Medizinhistorisches Journal* (2008) 43(3–4), pp. 327–43.

[19] Boudia, op. cit. (2), p. 72.

https://doi.org/10.1017/bjt.2026.10034 Published online by Cambridge University Press

'not unacceptable' risks to individuals and populations.[20] While higher doses meant more mutations, and thus a greater likelihood of cancer, even a single mutation posed a risk.

In the United States, partly because of the influential testimony of the NCI's Wilhelm Hueper, and the political influence of Representative James Delaney, Congress also famously distinguished carcinogenicity from other potential toxic effects through the so-called Delaney clause of 1958.[21] Unlike for other toxic effects, where it was possible to find a 'no observed adverse effect level' in an animal study and then multiply by a safety factor to determine a daily allowed dose for humans, the Delaney Clause mandated that there was never a safe dose for a carcinogen, whether that finding of carcinogenicity was made in animals or humans.[22] The Delaney clause had many complications from the start, not least of which were the exceptions carved out for substances like diethylstilbestrol (DES), allowing them to be used as long as no detectable amount was found in edible tissue.[23] Such special cases became increasingly common. As one of the FDA's lawyers described the paradox in retrospect, in some regards the Delaney clause had 'no force at all' because if an additive presented a significant risk, it was already not safe and therefore prohibited by statute, but if the cancer risk was so small as to be insignificant it could be 'read out of statute' by turning the issue from one of safety to one of detection.[24]

Nathan Mantel, who had continued working on extrapolating low-dose effects throughout the 1950s, found a new audience for his work after the 'great cranberry scare of 1959'.[25] A few weeks before Thanksgiving that year, Arthur Fleming, the Secretary of Health, Education, and Welfare (and therefore the person responsible for overseeing the FDA and NIH), issued a public warning that some of the year's cranberry crop was found to have remnants of aminotriazole, a chemical used in herbicides and linked to cancerous growth in lab rats. Empowered by the previous year's passage of the Food Additive Amendments, including the Delaney clause, Fleming's warning resulted in the restriction of sales of cranberries and essentially sank the $50 million-a-year industry right at its biggest moment. Though he rescinded the warning just days before Thanksgiving, noting that additional lots for sale had now tested 'clean' and could be safely consumed, the resulting furor and panic led an associate director of the NCI, G. Burroughs Mider, to turn to Mantel and his colleague, the long-time cancer researcher Ray Bryan, to develop protocols for reliably managing the problem of low-dose exposures.[26] Banning an entire food source because a very low level of a potentially harmful substance was detected had huge political consequences. As a result, Mantel and Bryan started from the assumption that 'we had to be willing to allow some level of risk'.[27] Rather than trying to define the specific risk of a substance, however, they worked

[20] Boudia, op. cit. (2), pp. 76–7.

[21] Mark Ryan Janzen, 'The cranberry scare of 1959: the beginning of the end of the Delaney clause', PhD dissertation, Texas A & M University, 2010, pp. 52–3; Charles H. Blank, 'The Delaney clause: technical naivete and scientific advocacy in the formulation of public health policies', *California Law Review* (1974) 62(4), pp. 1084–1120.

[22] Wargo, op. cit. (17), pp. 77–8, 121–2. Technically, the 'clause' as typically used refers to three parallel provisions for food additives (adopted 1958), color additives (1960) and animal drug residues (1968). Richard A. Merrill, 'FDA's implementation of the Delaney clause: repudiation of congressional choice or reasoned adaptation to scientific progress', *Yale Journal on Regulation* (1988) 5(1), pp. 1–88, 3.

[23] These complications included the exemption of 'foods generally recognized as safe' as well as the limitation to 'food additives', even though distinctions between 'raw' and 'processed' were not always obvious. Richard A. Merrill and Michael Schewel, 'FDA regulation of environmental contaminants of food', *Virginia Law Review* (1980) 66(8), pp. 1357–441, 1364; for more on the so-called DES proviso see Langston, op. cit. (17); Vogel, op. cit. (17).

[24] Merrill, op. cit. (22), p. 8.

[25] The phrase is from Michael Tortorello, 'The great cranberry scare of 1959', *New Yorker*, 24 November 2015, at www.newyorker.com/tech/annals-of-technology/the-great-cranberry-scare (accessed 20 February 2026); Tortorello heavily relies on Janzen, op. cit. (21).

[26] Mitchell H. Gail, 'A conversation with Nathan Mantel', *Statistical Science* (1997) 12(2), pp. 88–97, 91.

[27] Gail, op. cit. (26), p. 91.

https://doi.org/10.1017/bjt.2026.10034 Published online by Cambridge University Press

backwards, asking, if the regulative authority were willing to allow an additional risk of one in 100 million of developing cancer, for example, to what dose of a carcinogen should we allow exposure over a lifetime? They called this the 'virtual safe dose', acknowledging that what constituted safety and risk was inevitably in the eye of the beholder.

Published in 1961, Mantel and Bryan's paper defined 'virtual safe dose' by extrapolating downward from the upper limit on the 'true rate' of tumour incidence at a set confidence level. That is, once there is enough data (i.e. enough tumours appear in subject animals) to define *some* part of the dose–response curve, researchers can calculate a confidence interval at a given probability of error, take the upper limit of the interval, and then extrapolate downward to create a conservative model of what would happen at low doses. They used the modifier 'virtual' to emphasize that the actual effects at this dose would not have been observed experimentally. Basically, they argued for assuming the worst case given reliable data at moderate doses and extrapolating this to low doses in order to make what they hoped would be a conservative estimate of effects. Statistically, this wasn't all that sophisticated: what made the paper hugely influential was that it treated the statistical problem of low-dose extrapolation as a practical problem of clearly stating the assumptions by which regulators and researchers could make inferences from existing data.[28] Experiments with animals are inherently uncertain, especially at low doses when relatively few tumours are found, so Mantel and Bryan assumed that researchers will never be able to use animal studies to find an absolutely 'safe' dose for humans; the best anyone can do is to make a procedure by which one could determine a 'virtual' safe dose. Mantel and Bryan are clear that each of their choices could be made differently, but they are ultimately unavoidable specifications.[29]

Although, as Mantel's colleague noted, for years 'hardly anybody paid any attention to what they had said', by the end of the 1960s the work appeared to have new relevance as the FDA and later the EPA looked for new solutions to the problem of low-dose exposures to putative carcinogens.[30] As analytical methods of detection improved dramatically over the 1950s and 1960s, there were more and more additives found in the food supply, and more and more of them were found to cause cancer at a high enough doses, rendering them potentially subject to the Delaney clause if the substance was from additives or animal drugs.[31] The 1940s and early 1950s focus on bioassay for 'acute' toxicity – searching for short-acting, obvious effects – had yielded by the 1960s to a focus on long-term effects of low doses, particularly carcinogenic effects.

The head of Biometry and Epidemiology within the FDA's Bureau of Drugs was Charles Anello, who, when faced with the practical problems that low-dose extrapolation posed for FDA regulation, brought on his former teacher, Cornfield, to serve on the FDA's Panel on Carcinogenesis, part of the FDA Advisory Committee on Protocols for Safety Evaluation.[32] The resulting 1970 report noted a range of difficulties in trying to determine a safe dose of a potential carcinogen, as well as the inescapability of extrapolation. It was effectively

[28] Nathan Mantel and W. Ray Bryan, '"Safety" testing of carcinogenic agents', *Journal of the National Cancer Institute* (1961) 27(2), pp. 455–70.

[29] Mantel and Bryan, op. cit. (28), pp. 457–8.

[30] Testimony of Marvin Schneiderman, 'Risk/benefit analysis in the legislative process', *Hearing on Committee on Science and Technology Serial No. 71 before the Subcommittee on Science, Research and Technology (House) and Subcommittee on Science, Technology, and Space (Senate)*, Congress 96–1 (24–5 July 1979), p. 67.

[31] Wargo, op. cit. (17), pp. 107–10; Merrill, op. cit. (22), p. 14; Michael A. Grayson, 'A history of gas chromatography mass spectrometry (GC/MS)', in Michael L. Gross and Richard M. Caprioli (eds.), *The Encyclopedia of Mass Spectrometry*, Boston, MA: Elsevier, 2016.

[32] Food and Drug Administration Advisory Committee on Protocols for Safety Evaluation, 'Panel on carcinogenesis report on cancer testing in the safety evaluation of food additives and pesticides', *Toxicology and Applied Pharmacology* (1971) 20, pp. 419–38.

https://doi.org/10.1017/bjt.2026.10034 Published online by Cambridge University Press

impossible to show *non*-carcinogenicity with an experiment: if a test involved only a few animals and zero tumours resulted, a confidence interval centred at zero would still include the possibility of carcinogenicity; if a researcher made the sample size large enough to narrow that confidence interval, the chance of spontaneous tumour formation in an animal would rise, and in any case the confidence interval would *still* include *some* positive values. The bigger practical problem, however, was that multiple methods of extrapolation could be consistent with all existing experimental data and yet produce estimates of 'safe' doses that differed by multiple orders of magnitude.[33] Ultimately, the group pointed to the Mantel–Bryan procedure as one way to get an upper estimate, even if accuracy was not guaranteed. After the 1971 publication of the Panel on Carcinogenesis's recommendations, the FDA proposed a rule in 1973 specifying a 'practicable method' – i.e. Mantel–Bryan – for dealing with situations in which a putative carcinogen might enter the food supply.[34]

The importance of Mantel–Bryan for the FDA wasn't primarily that it gave reliable or statistically superior results. The citation, rather, was to the *approach* of Mantel–Bryan. If 'absolute safety can never be conclusively demonstrated experimentally', the best approach was to follow Mantel–Bryan in assigning an 'arbitrary but conservative level of maximum exposure resulting in a minimal probability of risk' for individuals.[35] In addition, procedures like Mantel–Bryan could be extended to incorporate data from multiple studies, synthesizing a range of dosing experiments into one extrapolation procedure. In 1976, the EPA issued its first 'guidelines' for establishing carcinogens (the Office of Pesticides was transferred from the USDA to the EPA upon the latter's creation), which promoted epidemiological studies as the 'best evidence' but noted that in practice they had to be combined with 'risk extrapolation procedures' such as Mantel–Bryan.[36]

By the mid-1970s, models of low-dose extrapolation had become critical in estimating carcinogenic risks.[37] In 1960, for example, feed contaminated with aflatoxin, a mould commonly found on peanuts and corn, had been linked to the dramatic deaths of otherwise healthy turkeys in the United Kingdom. Soon, it became clear from laboratory tests that aflatoxins were carcinogenic in many animals, and it seemed in humans too, based on epidemiological studies in Asia and Africa linking liver cancer to high levels of peanut consumption. (Peanuts, like some other food products, host aflatoxins, a form of mould-produced toxins; the fact that they are 'naturally occurring' meant that the Delaney clause was not applicable so an explicit tolerance level needed to be established.) International committees like the World Health Organization's Protein Advisory Group were using dose–response experiments on monkeys to try to set a 'safe' level of peanut consumption in member countries.[38] When the FDA turned in 1978 to potential regulation of aflatoxin-infected peanuts, they needed a method for combining evidence from epidemiological and laboratory data and estimating risks to health. Epidemiological studies had suggested the lifetime liver cancer incidence rates were 161 per 100,000 for the entire US, but these observational studies had provided little information about the risks of specific 'doses' of peanuts, let alone clear-cut evidence that aflatoxins are carcinogenic in humans. The approach the

[33] David G. Hoel, David W. Gaylor, Ruth L. Kirschstein, Umberto Saffiotti and Marvin A. Schneiderman, 'Estimation of risks of irreversible, delayed toxicity', *Journal of Toxicology and Environmental Health* (1975) 1(1), pp. 133–51, 142.

[34] 'Compounds used in food-producing animals', *38 Federal Register*, 1973, pp. 19226–30, 19226.

[35] 'Compounds used in food-producing animals', op. cit. (34), p. 19227.

[36] Environmental Protection Agency, *Interim Procedures and Guidelines for Health Risk and Economic Impact Assessments of Suspected Carcinogens* (1976), Appendix I, p. 5.

[37] David Demortain, *The Science of Bureaucracy: Risk Decision-Making and the US Environmental Protection Agency*, Cambridge, MA: MIT Press, 2020, 56.

[38] Lucas M. Mueller, 'Risk on the negotiating table: malnutrition, mold toxicity, and postcolonial development', in Creager and Gaudillière, op. cit. (17).

https://doi.org/10.1017/bjt.2026.10034 Published online by Cambridge University Press

FDA took was to estimate lifetime liver cancer incidence rates from five studies and use a revised version of the Mantel–Bryan procedure to combine these studies into one estimate of risk. That this estimate was ultimately higher than the overall incidence rates was a good sign that the extrapolations were conservative. Then the studies had to estimate the effects of lowering the allowable tolerances for aflatoxins in terms of both cancer incidence and the costs to food producers, while remaining cognizant of detection tolerances.[39]

Through the late 1970s and into the 1980s there were significant legal challenges to the use of extrapolation methods rather than tolerance thresholds but nothing superseded them, in part because, as the FDA's acting commissioner Sherwin Gardner explained in 1977, methods like Mantel–Bryan had value in providing a 'rational, uniform procedure' that has neither the problem of limiting to 'practical zero' residue nor that of limiting to the 'lowest detectable' amount.[40] Statistical methods were not praised as accurate, per se – different mathematically valid and experimentally calibrated methods could and did give estimates of low-dose effects that differed by multiple orders of magnitude; judgement was required as to which ones might be best to apply.[41] Rather, regulators praised certain methods like Mantel–Bryan for their 'rationality' and 'uniformity'. In effect, FDA regulators made it clear that their goal was not to accurately predict dose–response effects, but to act conservatively, given the inescapable uncertainty.

Statisticians as sceptical quantifiers

Multiple models for estimating responses to low doses of exposure continued to persist for decades after the 1960s because there was little biological evidence for choosing one of them over another. Statisticians themselves noted by 1974 that it was unlikely that any statistical procedure would be able to accurately model cancer.[42] Indeed, in 1983, the FDA contracted with the National Research Council to produce *Risk Assessment in the Federal Government: Managing the Process*. Noting that there was 'no ready solution', the NRC admitted that the 'dominant analytic difficulty is pervasive uncertainty'.[43] Nevertheless, in the case of dose–response extrapolation, the solution was to use multiple models, and because it was 'impossible to distinguish their validity on the basis of goodness of fit', they recommended considering the 'biological plausibility' of each model.[44] At no point did one statistical model of low-dose extrapolation rise to the level of general acceptance, and, as one commentator noted dismissively, there is usually 'no reliable experimental basis for selecting one extrapolative model over another'.[45] Even the assumption of a stark difference between acute and reversible as opposed to cumulative and irreversible effects did not hold up to biological or statistical scrutiny.[46] It was obvious by the 1980s that the statistical problem of modelling low-dose carcinogenicity remained (in the assessment of one of its most

[39] Food and Drug Administration, Bureau of Foods, *Assessment of Estimated Risk Resulting from Aflatoxins in Consumer Peanut Products and Other Food Commodities* (19 January 1978).

[40] 'Chemical compounds in food-producing animals', *42 Federal Register*, 22 February 1977, pp. 10412–37, 10419.

[41] Food and Drug Administration Advisory Committee on Protocols for Safety Evaluation, op. cit. (32), p. 433.

[42] Nanak Chand and D.G. Hoel, 'A comparison of models for determining safe levels of environmental agents', in Frank Proschan and R.J. Serfling (eds.), *Reliability and Biometry: Statistical Analysis of Lifelength*, Philadelphia: Society for Industrial and Applied Mathematics, 1974, 698.

[43] National Research Council (US), *Risk Assessment in the Federal Government: Managing the Process*, Washington, DC: National Academy Press, 1983, pp. 6, 11.

[44] National Research Council, op. cit. (43), pp. 25–7.

[45] Howard Latin, 'Good science, bad regulation, and toxic risk assessment', *Yale Journal on Regulation* (1988) 5(1), p. 99.

[46] Jerome Cornfield, 'Carcinogenic risk assessment', *Science* (1977) 198(4318), pp. 693–9.

https://doi.org/10.1017/bjt.2026.10034 Published online by Cambridge University Press

distinguished practitioners, Peter Armitage) 'extremely difficult' despite a 'voluminous' literature over the previous three decades.[47]

Because dose–response calculations lay at the heart of most methods of estimating risk, government statisticians were also sceptical of the move to what has come to be called quantitative risk analysis. Dose–response methods are sometimes framed as part of a prehistory of what has been coined the 'economic style of reasoning', in which quantitative risk analysis, cost–benefit analysis and other economic approaches have become pervasive in public policy and everyday life.[48] Wargo, for example, noted that the FDA's 1973 specification of Mantel–Bryan as a 'practical method' for dealing with potentially carcinogenic substances was 'a watershed in the history of U.S. attempts to regulate environmental carcinogens because it ties the prohibitive power of the Delaney clause to quantitative risk assessment methods'.[49] If different pesticide tolerance levels, for example, were estimated to have different effects on cancer incidence, then one had to balance that with the effect that tightening those tolerances would have on the food supply. It was a balance in which regulators had to calculate, for example, 'how many cancers to permit in order to assure the availability of food at a reasonable price'.[50] The state of the art was encapsulated in the 1979 report of the Interagency Regulatory Liaison Group (IRLG), 'Scientific bases for identification of potential carcinogens and estimation of risks'.[51] Written specifically by the Work Group on Risk Assessment, which included representatives from a range of federal agencies, as well as statistician Marvin Schneiderman from the Cancer Institute, the report laid out methods after a decade of rapid developments. Even if not regulatory or statutory policy per se (the agencies themselves never adopted the guidelines and increasingly went separate ways), the report clearly indicated that the field was moving to some kind of quantitative risk analysis, and ultimately to a cost–benefit calculation.[52] Boudia's account of the 1983 NRC report likewise notes that the combination of a no-threshold model for cancer and the impossibility of banning most putative carcinogens led to the introduction of risk assessment as an alternative framework, one consistent with the formal use of statistical procedures.[53]

Perhaps ironically, statisticians doing the low-dose extrapolations were themselves far more circumspect about risk assessment in general and risk–benefit analysis in particular. Umberto Saffioti, an NCI scientist who would work closely with the EPA in designing their guidelines in the 1970s, warned about the problems of false precision and downplaying uncertainty, going so far as to oppose any discussion of quantitative risk assessment methodologies.[54] NIH statistician Samuel Greenhouse noted in congressional testimony in

[47] P. Armitage, 'The assessment of low-dose carcinogenicity', *Biometrics* (1982) 38, pp. 119–29, 119.

[48] Elizabeth Popp Berman, *Thinking like an Economist: How Efficiency Replaced Equality in U.S. Public Policy*, Princeton, NJ: Princeton University Press, 2022, p. 3; Michelle Murphy, *The Economization of Life*, Durham, NC: Duke University Press, 2017.

[49] Wargo, op. cit. (17), p. 111.

[50] Merrill and Schewel, op. cit. (23), p. 1426.

[51] Interagency Regulatory Liaison Group, Work Group on Risk Assessment, 'Scientific bases for identification of potential carcinogens and estimation of risks', *Journal of the National Cancer Institute* (1979) 63(1), pp. 241–68.

[52] Demortain, op. cit. (37), p. 75. The grounding of risk assessment in nuclear regulation is traced in Ellen Bales, 'An element of uncertainty: radon and the quantification of risk in 20th-century America, 1945–2000', PhD dissertation, University of California, Berkeley, 2009; Kimberly M. Thompson, Paul F. Deisler Jr and Richard C. Schwing, 'Interdisciplinary vision: the first 25 years of the Society for Risk Analysis (SRA), 1980–2005', *Risk Analysis* (2005) 25(6), pp. 1333–86; National Research Council, op. cit. (43); Demortain, op. cit. (37), p. 75.

[53] Soraya Boudia, 'Managing scientific and political uncertainty: environmental risk assessment in historical perspective', in Boudia and Jas, op. cit. (17), p. 107.

[54] Demortain, op. cit. (37), p. 70; Marc Karnis Landy, Marc J. Roberts and Stephen R. Thomas, *The Environmental Protection Agency: Asking the Wrong Questions*, Oxford: Oxford University Press, 1990, 193.

https://doi.org/10.1017/bjt.2026.10034 Published online by Cambridge University Press

1979 that the 'logistical difficulties' in conducting risk–benefit analysis for every substance would be 'tremendous' and overwhelming.[55] In his own work at the National Institute of Child Health and Human Development, Greenhouse had to do such analyses, and found that the assumptions required not just 'gross approximations' but also tricky translations from risks to dollar figures. He recommended continuing to use a 'qualitative' approach. At the same hearing, the NCI's Schneiderman had a more technical qualm, namely that the methods of risk computation were never foolproof, and we didn't know which methods had the best biological bases or approach.[56] Two years later, Schneiderman was called to testify on the use of cost–benefit analyses in the context of the EPA's establishment of air quality standards. Schneiderman noted not only that risk calculations are sensitive to the models chosen in the case of low-dose extrapolations, but also that the entire edifice of cost–benefit analysis is still an 'infant field', more an 'art form' than a science.[57] His prepared statement went further, noting that any cost–benefit analysis 'collapses all the relevant factors' into 'a single number', 'thus suspending human thought and judgment and important, unmeasurable things'.[58] By hiding behind often arbitrary yet decisive technical choices like statistical models and discount rates, the procedure fails to account for human judgement, particularly about things that happen not to be easily measurable.

Schneiderman and his colleagues rejected quantification for quantification's sake; the statisticians' goal was not to turn everything into numbers. They recognized the way in which industry might fudge numbers, and indeed risk–benefit analysis was eventually used in place of risk-only calculations in the vast majority of cases, which is precisely what industry representatives wanted, aware as they were of the generative power of uncertainty.[59] Statisticians' goal, rather, was to marshal evidence in an orderly fashion, with assumptions outlined and uncertainty highlighted. The essential role of statistics, according to the NIH statisticians, was to facilitate judgements by establishing 'reasonable, objective' procedures for making them.[60] Importantly, objectivity here was a descriptor for the *process*, not the *product*. A 'reasonable, objective' procedure did not reliably yield accurate estimates; it was simply the best method to use. In contrast, methods that collapsed relevantly distinct categories, deployed misleading approximations, or that hid the judgements and guesses of investigators, could not be objective in this sense. Though some degree of arbitrariness in making judgements was unavoidable, NIH statisticians emphasized the need for procedures to be explicitly linked to existing experimental evidence, ideally high-quality experiments. Statisticians did not present one method to regulators as a uniquely valid way of making decisions: many different methods were valid, statistically speaking, and all of them required multiple assumptions to be made about extrapolation procedures and biological mechanisms. Their argument for adopting statistical measures was that the best methods 'facilitated judgments', making regulators' (inevitable) assumptions and simplifications visible and accountable.

[55] Testimony of Samuel Greenhouse in 'Risk/Benefit Analysis in the Legislative Process', op. cit. (30), pp. 70–71.

[56] Schneiderman, op. cit. (30), p. 67.

[57] Testimony of Marvin Schneiderman, 'Health standards for air pollutants', *Hearing on Committee on Energy and Commerce Serial No. 97-97 before the Committee on Energy and Commerce: Subcommittee on Health and the Environment*, Congress: House 97–1 (14 October 1981), pp. 22–3.

[58] Schneiderman, op. cit. (57), p. 27.

[59] Wargo, op. cit. (17), pp. 125–6; more broadly, see Naomi Oreskes and Erik M. Conway, *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming*, New York: Bloomsbury Press, 2011.

[60] Nathan Mantel and Marvin A. Schneiderman, 'Estimating "safe" levels, a hazardous undertaking', *Cancer Research* (1975) 35(6), pp. 1379–86, 1386.

https://doi.org/10.1017/bjt.2026.10034 Published online by Cambridge University Press

Statistics within clinical trials

Though the use of statistics for designing and interpreting clinical trials is a much better-known case than their role in dose–response models, it is only in hindsight that the introduction of statistical methods into new drug approval procedures appears so successful. In the 1950s, formal clinical trials were often seen as overly complicated and expensive, and the mechanisms of blinding and randomization seemed counterintuitive to physicians used to treating individual patients in all their specificity. Statistics were fine for aggregate public-health indicators – the incidence and prevalence of disease or different measures of mortality – but seemingly had little relevance for clinicians. As late as 1944, a discussion of how to evaluate new drugs in the *Journal of the American Medical Association* did not envision any role for statistically validated randomized controlled trials.[61] Nevertheless, by the 1970s trial results were essentially mandatory for new drug applications, and by the 1980s reformers were focused on measures for improving the quality and process of the trials.[62] The basic outlines of the story are well known – the first highly publicized statistically interpreted clinical trials emerged in the late 1940s, leading gradually to their establishment as a 'gold standard' for clinical knowledge by the 1970s.[63] The problem is that such an outline underplays the problems of trials, while also not explaining the tensions among statisticians as to what trials could show, mathematically.

Though informal trials and other tests of drugs' safety and efficacy had long been submitted, trials only formally became a part of the FDA's rule-making in the wake of the so-called 1962 Kefauver-Harris Amendments (to the Federal Food, Drug, and Cosmetic Act), which expanded and clarified the authority of the FDA in the regulation of pharmaceuticals.[64] Even in the 1960s it was not obvious that formal, statistically interpreted clinical trials would be required. Drug companies had long relied upon informal 'testing' by conducting clinical 'trials' with free samples distributed within networks of friendly physicians, with the resulting experience of clinicians reported back and aggregated. The question was not whether formal statistically interpreted clinical trials *could* be used as evidence, but whether they *had* to be used *exclusively*.[65] By 1973, the FDA's rules specifying which types of 'controlled comparisons' would allow 'quantitative evaluation' had survived legal challenges from pharmaceutical firms, and the US Supreme Court affirmed the agency's authority in setting the standards for what counted as adequate evidence.[66]

FDA regulators were not starting from scratch in figuring out the role of statistics in clinical trials. They had close connections with statistical colleagues at the NIH (as well as with the pioneering Medical Research Council in the UK), and as one of the primary people charged with implementing the 1962 amendments at the FDA would later remember, 'some of the people at NIH, they knew all about how to do a good study in the 1950's, but it took us a

[61] Walton van Winkle, 'Laboratory and clinical appraisal of new drugs', *Journal of the American Medical Association* (1944) 126(15), pp. 958–61.

[62] Thomas C. Chalmers, H. Smith Jr, B. Blackburn, B. Silverman, B. Schroeder, D. Reitman and A. Ambroz, 'A method for assessing the quality of a randomized control trial', *Controlled Clinical Trials* (1981) 2(1), pp. 31–49.

[63] Marks, op. cit. (1); Laura Bothwell, 'The emergence of the randomized controlled trial: origins to 1980', PhD dissertation, Columbia University, 2014.

[64] Bothwell, op. cit. (63); Suzanne White Junod, 'FDA and clinical drug trials: a short history', *About FDA*, 7 July 2014, at http://web.archive.org/web/20160323021459/http://www.fda.gov/AboutFDA/WhatWeDo/History/Overviews/ucm304485.htm (accessed 18 February 2026); Marks, op. cit. (1); Carpenter, op. cit. (2).

[65] Charles Anello, 'FDA principles on clinical investigations', *FDA Papers* (1970) 4(5), pp. 14–15, 23–4; Anello and Junod, op. cit. (12); Division of Medical Sciences, National Research Council, *Drug Efficacy Study: Final Report to the Commissioner of Food and Drugs*, FDA contract nos. 66–197, Washington, DC: National Academy of Sciences, 1969.

[66] 21 CFR (1971), §130.12(a); Peter Barton Hutt, 'Views on supreme court/FDA decisions', *Food Drug Cosmetic Law Journal* (1973) 28(11), pp. 662–75.

https://doi.org/10.1017/bjt.2026.10034 Published online by Cambridge University Press

long time to learn all about it'.[67] The experience with trials at the NIH was indeed expansive, and while many in the 1950s were tied to the Clinical Center and the Cancer Institute, where many of the NIH's statisticians were located, there had been other highly visible examples in that period where the usefulness of trials was firmly established, including the study of the effects of oxygen on retrolental fibroplasia in newborns and the trial of Jonas Salk's killed-virus vaccine for poliomyelitis.[68]

The statistical machinery for these trials was never 'off the shelf'; it was always custom-built for the particular challenges of the situation. The net result was that the statistical concepts could be used to directly address the difficulties posed by 'hard cases' in which harms and benefits had to be carefully weighed and balanced. Because those harms and benefits were usually unknown at the outset, these cases helped to demonstrate the superiority of formal randomized clinical trials over aggregated reports from physicians.

There were many reasons for the introduction of statistically interpreted trials into regulatory decision making. There was, as the foremost historian of medical reform in this period explained, the crucial role that they played in the mid-century battle between 'clinical expertise and physician autonomy' as set against 'bureaucratic dictates'.[69] Statisticians in this way might have participated in the mid-century erosion of physician authority and autonomy.[70] And as another history of clinical trials has argued, trials turn a complex social question (does this drug work?) into a technical procedure backed by the authority of statistical theory.[71] More generally, numbers have long been a mode by which those with less authority (e.g. regulators as compared to mid-century physicians) might intrude in the decision-making process by suggesting that only numbers enable objective judgements to be made.[72]

As convincing as these explanations are, statistically interpreted trials remained an awkward fit with the goal of determining whether a new drug was safe or effective for specific patients. Most trials, some statistician–critics noted, were designed using frequentist models of statistics, and therefore not really about 'degrees of belief' in a single proposition so much as about what was likely to happen in the 'long run'. (And what constituted the 'long run' was unclear when the goal was to assess the carcinogenicity or efficacy of a specific treatment in the present.) Even at the height of the promotion of trials in the 1960s, NIH statisticians publicly questioned whether statistical methods really provided the kind of evidence regulators and researchers were asking of them. Questions of dosing and proper outcome measurements, complications from non-compliance or discrepancies

[67] Peter Barton Hutt and Robert Temple, 'Commemorating the 50th anniversary of the drug amendments of 1962', *Food and Drug Law Journal* (2013) 68(4), pp. 449–66, 459; Anello and Junod, op. cit. (12), p. 262.

[68] On retrolental fibroplasia, the statistical details are discussed in Samuel W. Greenhouse, 'Some historical and methodological developments in early clinical trials at the National Institutes of Health', *Statistics in Medicine* (1990) 9(8), pp. 893–901; Bothwell, op. cit. (63), puts the trial in the context of clinical trials; William A. Silverman, *Retrolental Fibroplasia: A Modern Parable*, New York: Grune & Stratton, 1980, situates the trial in the context of RLF more broadly. And on polio see Harry M. Marks, 'The 1954 Salk poliomyelitis vaccine field trial', *Clinical Trials* (2011) 8(2), pp. 224–34; David M. Oshinsky, *Polio: An American Story*, Oxford: Oxford University Press, 2005; Liza Dawson, 'The Salk polio vaccine trial of 1954: risks, randomization and public involvement in research', *Clinical Trials* (2004) 1(1), pp. 122–30; Marcia Meldrum, '"A calculated risk": the Salk polio vaccine field trials of 1954', *British Medical Journal* (1988) 317(7167), pp. 1233–6; Paul Meier, 'Polio trial: an early efficient clinical trial', *Statistics in Medicine* (1990) 9(1–2), pp. 13–16.

[69] Marks, op. cit. (1), p. 225.

[70] Keith Wailoo, 'Sovereignty and science: revisiting the role of science in the construction and erosion of medical dominance', *Journal of Health Politics, Policy and Law* (2004) 29(4–5), pp. 643–60.

[71] Marcia Lynn Meldrum, '"Departures from the design": the randomized clinical trial in historical context, 1946–1970', PhD dissertation, State University of New York at Stony Brook, 1994.

[72] Theodore M. Porter, *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*, Princeton, NJ: Princeton University Press, 1996.

https://doi.org/10.1017/bjt.2026.10034 Published online by Cambridge University Press

among multiple centres, the role of 'early looks' at the data and the resulting messiness of *p*-values, all ensured that only rarely did a trial result in a clear-cut answer about safety and efficacy.[73] Trials required controlled settings and consequently carefully selected and monitored patients, a group often quite different from the population for which the drugs would eventually be prescribed. There were always reasons why a trial of a particular drug would not adequately mimic physicians' actual practices in the clinic, rendering any inferences from the trial problematic for practising physicians.[74] Such criticisms also made the authority of statistically interpreted trials relatively easy to undermine: when clinical trials or toxicity studies challenged widely held presumptions about commonly used oral hypoglycemics or artificial sweeteners, for example, industry advocates and powerful physicians were able to step in and argue for exceptions to be made and statistical findings to be largely overruled.[75]

Like the case of dose–response extrapolation methods, the statistics of clinical trials were clearly useful for regulators, but even the government statisticians promoting them did not think they provided a foolproof method for ascertaining truths about the effectiveness and safety of drugs. Rather, they provided a way to make assumptions underlying the calculations and conclusions of the approval process explicit. They served as a practical, 'rational' approach to decision making, even if there were heated disputes about the accuracy of the conclusions. Statistical methods did not eliminate subjective expertise or individual judgements but made those decisions legible to regulators by requiring specific evidence to be supplied for them rather than allowing them to be made by hand-waving deference to expertise or informal experience.

Conclusions

The statisticians most central to the development of statistics for federal regulators were clearly not eager to turn everything into numbers. They were not, that is, like the paradigmatic case of army engineers or accountants famously introduced by Theodore Porter in *Trust in Numbers* as emblematic of the twentieth-century bureaucratic move to quantify.[76] Judgement here was not to be avoided but to be made explicit. Variability and uncertainty were not reducible in most cases, but they could be bounded enough to move forward with practical regulations. The language of objectivity by Mantel, Schneiderman and other statisticians was not directed to the conclusions that regulators made – valid statistical methods could and did yield wildly divergent estimates of effects – but was made in contrast to methods that ignored assumptions, hid investigators' judgement calls and collapsed important distinctions in the data. The best statistical models for regulators were ones that recognized the need for the judgements and assumptions lurking behind quantitative models to be made explicit and justified.[77]

[73] For one such discussion at the NIH see S.J. Cutler, S.W. Greenhouse, J. Cornfield and M.A. Schneiderman, 'The role of hypothesis testing in clinical trials', *Journal of Chronic Diseases* (1966) 19, pp. 857–82.

[74] The paradigmatic case for this kind of criticism in the 1970s was that of the University Group Diabetes Project; see relevant discussions in Marks, op. cit. (1); and Jeremy A. Greene, *Prescribing by Numbers: Drugs and the Definition of Disease*, Baltimore: Johns Hopkins University Press, 2007.

[75] Greene, op. cit. (74), pp. 115–47; Richard A. Merrill and Michael R. Taylor, 'Saccharin: a case study of government regulation of environmental carcinogens', *Virginia Journal of Natural Resources Law* (1985) 5(1), pp. 1–84.

[76] Porter, op. cit. (72); Theodore M. Porter, 'Quantification and the accounting ideal in science', *Social Studies of Science* (1992) 22, pp. 633–52.

[77] This is consistent with a broader mid-century attempt to frame statistical methods not as a mode of objective fact finding, but rather as the best available process for making inescapably subjective judgements reliable and legible. Christopher J. Phillips, 'The taste machine: sense, subjectivity, and statistics in the California wine world', *Social Studies of Science* (2016) 46(3), pp. 461–81.

https://doi.org/10.1017/bjt.2026.10034 Published online by Cambridge University Press

The example of statistics in regulation provides a contrast with Berman's account of the rise of 'economic style of reasoning' in US public policy. Berman's *Thinking like an Economist* centres on the rise of market-based, trade-off-focused and efficiency-minded rule-making on both the left and the right in the 1980s. As she rightly notes, the initial flurry of social policy surrounding health and the environment around 1970 was not a good example of the 'economic reasoning' that would eventually take hold.[78] Nevertheless, she charts how the informal use of balancing risks and benefits in the 1970s transitioned to formal requirements for cost–benefit analysis in the 1980s as the 'economic style of reasoning' came to dominate public policymaking. For the economists that she focused on this is a triumph, but the statisticians – those by and large creating the procedures and inferential methods that underlay the calculations – were rarely so sanguine. Models that collapsed relevant distinctions and hid investigators' assumptions in the name of producing a number that could be taken as authoritative were suspect. Moreover, economists' models were both descriptive and active, providing supposedly accurate depictions of how markets and economies really functioned, as well as serving as a force that transformed those very markets.[79] The statisticians who worked on dose–response models were always clear that their models may or may not bear much relation to underlying biological processes and that consequently inferences from them should be made cautiously. Statisticians' methods emphasized uncertainty and variability, not the power of prediction. The triumph of economics was a loss for statistics.

Statistics for regulatory decision making was of a piece with statisticians' emphasis on being collaborators in a complex world of often conflicting and ambiguous scientific evidence. Statisticians in this period served prominently on high-profile 'blue-ribbon' panels such as the Surgeon General's report on smoking and health (William Cochran) and the President's Science Advisory Committee Panel on Environmental Pollution (John Tukey). On these committees, statisticians played decisive roles in assessing scientific data and facilitating consensus. Cochran was responsible, for example, for synthesizing the range of epidemiological studies in order to establish the causal association between cigarette smoking and lung cancer that was at the centre of the Surgeon General's report, while Tukey spoke repeatedly about the need for judgement, and about the need to avoid any mechanization of decision making.[80] They had the contemporary model of their close collaborators at the NIH, including Cornfield, Mantel, Schneiderman and Greenhouse, who provided similar collaborative expertise to the FDA. In testimony before Congress, Greenhouse noted that in his

> experience as a consultant to the FDA, I have found these decisions to be very complex. They are usually so because the scientific component of the evidence is rarely cut and dried, as lay individuals would prefer to have it. In most cases, studies and the data they yield contain some flaw, some ambiguity which make the evidence far

[78] Berman, op. cit. (48), p. 159.

[79] Donald MacKenzie, *An Engine, Not a Camera: How Financial Models Shape Markets*, Cambridge, MA: MIT Press, 2006.

[80] Allan M. Brandt, *The Cigarette Century: The Rise, Fall, and Deadly Persistence of the Product That Defined America*, New York: Basic Books, 2007, p. 222; Charles A. LeMaistre, Donald R. Shopland Sr, Emmanuel Farber, Eugene H. Guthrie and Peter V.V. Hamil, *Clearing the Air: The Untold Story of the 1964 Report on Smoking and Health*, San Francisco: University of California Health Humanities Press, 2024, pp. 200–1; Advisory Committee to the Surgeon General of the Public Health Service, *Smoking and Health*, Public Health Service Publication No. 1103, Washington, DC: US Department of Health, Education, and Welfare, 1964; John W. Tukey, 'What have statisticians been forgetting?', in *The Collected Works of John W. Tukey: Philosophy and Principles of Data Analysis, 1965–1986* (ed. Lyle V. Jones), 4 vols., Monterey: Wadsworth & Brooks, 1986, vol. 4, pp. 587–99; Alexander Campolo, '"Thinking, judging, noticing, feeling": John W. Tukey against the mechanization of inferential knowledge', *KNOW: A Journal on the Formation of Knowledge* (2021) 5(1), pp. 83–111.

https://doi.org/10.1017/bjt.2026.10034 Published online by Cambridge University Press

> from conclusive … It then clearly becomes a matter of judgment on the part of the regulatory agency …[81]

Statistical methods were never a satisfactory machine for making reliable decisions. They were useful tools for helping to forge a consensus when regulatory action was necessary, but experimental data remained messy, uncertain and necessarily incomplete.

That is not to say that the consensus-making efforts of statisticians somehow reliably 'worked'. Nearly contemporary research by STS scholars like Sheila Jasanoff noted high-profile 'flawed decisions' by regulators in the late 1970s that led critics to highlight the failures of scientific 'peer review'. Many of the motivating examples in her 1990 book *The Fifth Branch* were focused on the problems of calculating risk from low-dose exposures (such as nitrites added to food, the pesticide 2,4,5-T on food crops and occupational exposure to toxins).[82] Though the focus of the book was on the use of science by federal regulatory agencies like the EPA and FDA, examples like these actually turned on the question of how to extrapolate risk from low-dose exposures. One controversy she examines, namely the calculation of what percentage of cancer cases are attributable to occupational exposure, is presented as a choice between flawed science and politically motivated science. While it may well have been the case that some estimates of occupational sources of cancer were wildly exaggerated to score political points (and that some statisticians were more technically sophisticated than others), the scientific debate is better understood as a methodological debate about, among other things, whether one should work forward from estimates of relative risk to population effects or back from epidemiological data to apportion deaths to various causes. (Ironically, one of the key players in this episode was Schneiderman, the NCI statistician and consistent sceptic of simplified risk calculations who was cited as part of the authority behind the risk calculations made from doses of exposure.) From another point of view, the debate was perhaps an inevitable by-product of the difficulty of making estimates of risk from occupational exposures. Jasanoff's analysis that episodes like this challenged regulators' authority is absolutely right, but the methods themselves were inconclusive precisely because estimates of the effects of particular doses of exposure were so uncertain. As the FDA, EPA and other agencies increasingly relied on statistical methods in this period, their use remained contested, even among statisticians.[83]

From a remove of over four decades, it is now clear that the resolution of such contests is not just a matter of awaiting good-enough science to create 'facts' about risk or about how such science might be used or misused by regulators.[84] After all, debates about 'low-dose' exposures were *never* satisfactorily resolved. The phrase 'low-dose toxicity' initially succeeded in distinguishing dose–response research from the discredited model of threshold doses, but scholars have tracked how debates and controversies continued long after the 1980s as 'low-dose' exposures became a 'necessary evil' that needed to be managed, while

[81] 'FDA's regulation of the drug "Triazure"', *Hearing before the Committee on Government Operations*, US House of Representatives 94th Congress, 2nd Session (27 October 1976), 17.

[82] Jasanoff, *The Fifth Branch*, op. cit. (3), pp. 20–32.

[83] Jasanoff, op. cit. (3), pp. 199–200; a contemporary overview of the issues, on which Jasanoff relies, is R. Jeffrey Smith, 'Government says cancer rate is increasing: but are the data really there? And are chemicals really the cause?', *Science* (1980) 209(4460), pp. 998–1000; the most influential statistical account of how best to calculate such figures was R. Doll and R. Peto, 'The causes of cancer: quantitative estimates of avoidable risks of cancer in the United States today', *Journal of the National Cancer Institute* (1981) 66(6), pp. 1191–308; on the 1981 article's long influence (and accuracy) see W.J. Blot and R.E. Tarone, 'Doll and Peto's quantitative estimates of cancer risks: holding generally true for 35 years', *Journal of the National Cancer Institute* (2015) 107(4), pp. 1–5.

[84] Cf. Jasanoff, op. cit. (3), p. 216.

https://doi.org/10.1017/bjt.2026.10034 Published online by Cambridge University Press

experiments remained difficult to replicate and uncertainties remained pervasive.[85] This literature has focused on the public-policy implications of low-dose extrapolations rather than on the statistical methods and choices at the heart of these claims. The government statisticians who were developing the methods enabling such extrapolations in the 1960s and 1970s were never convinced that their methods were anything other than convenient statistical fictions – hence the phrasing of 'virtual safe doses' – even as they promoted the methods as necessary for making regulators' judgements explicit.

In the end, only in disciplines *outside* statistics proper (psychology, economics and education, above all), were simple algorithmic uses of statistics deemed acceptable for the creation of reliable knowledge at mid-century.[86] Statisticians themselves were typically not so sanguine, and certainly those involved in FDA rule-making had few illusions about the ability of statistics to provide answers to complex regulatory questions. Moreover, the precise period of the spread of statistics in regulatory agencies – the long 1960s – was also that of a fundamental crisis within statistical theory as to the philosophical basis for statistical inference. Indeed, during the same year as the influential 1962 amendments on drug approvals passed, statisticians were vehemently debating the entire foundation for making judgements about data.[87] These were also the years in which statistics departments were being proposed, and the place of the field within the academy was very much in flux.[88] Despite the growing reliance of regulators on statistical methods in this period, statistical theory was not monolithic. By taking seriously debates *among* statisticians and seeing statistics as a dynamic and contested field rather than a set of static tools, the work that statistics could and could not do for regulators becomes clearer.

Statisticians promoted tools to make rational, specifiable procedures for generating regulations in cases where evidence was conflicting, results were variable and uncertainty was rife. Statisticians who helped regulators at the FDA and beyond portrayed themselves as professional consultants, called upon to make sense of the evidence because *some* decision had to be made. They were the experts, not in reducing judgement to rote calculation, but in designing studies and aggregating data and evidence so that judgements could be effectively deployed in the service of regulating health and medicine.

Acknowledgements. Research for this article was supported by the National Library of Medicine of the National Institutes of Health under Award Number G13LM013556 and a Michael E. Debakey Fellowship in the History of Medicine. The content is solely the responsibility of the author and does not necessarily represent the official views of the National Institutes of Health. Thanks also to two anonymous referees, to Joel Greenhouse and especially to Lara Keuck, Angela Creager and the other participants in the Validation and Regulation in the Sciences of Health online and in-person workshops.

Competing interests. The authors declares none.

[85] Boudia, op. cit. (2); Sheldon Krimsky, 'Low-dose toxicology: narratives from the science–transcience interface', in Boudia and Jas, op. cit. (17).

[86] Christopher J. Phillips, 'Inference rituals: algorithms and the history of statistics', in Morgan G. Ames and Massimo Mazzotti (eds.), *Algorithmic Modernity*, Oxford: Oxford University Press, 2023.

[87] Allan Birnbaum, 'On the foundations of statistical inference', *Journal of the American Statistical Association* (1962) 57(298), pp. 269–306; L.J. Savage et al., 'On the foundations of statistical inference: discussion', *Journal of the American Statistical Association* (1962) 57(298), pp. 307–26.

[88] The 1960s were a key period for many statistics departments: Alan Agresti and Xiao-Li Meng (eds.), *Strength in Numbers: The Rising of Academic Statistics Departments in the U.S.*, New York: Springer, 2013.

Cite this article: Christopher J. Phillips, 'The history of statistics in statistically valid regulation,' *BJHS Themes* (2026), pp. 1–17. https://doi.org/10.1017/bjt.2026.10034

https://doi.org/10.1017/bjt.2026.10034 Published online by Cambridge University Press