It Is Time to Reconsider Factorial Designs: How Bradford Hill and R. A. Fisher Shaped the Standard of Clinical Evidence

Duncan Neuhauser, PhD; Shannon M. Provost, PhD; Lloyd P. Provost, MS

Objectives: Could medical research and quality improvement studies be more productive with greater use of multifactor study designs? **Methods:** Drawing on new primary sources and the literature, we examine the roles of A. Bradford Hill and Ronald A. Fisher in introducing the design of experiments in medicine. **Results:** Hill did not create the randomized controlled trial, but he popularized the idea. His choice to set aside Fisher's advanced study designs shaped the development of clinical research and helped the single-treatment trial to become a methodological standard. **Conclusions:** Multifactor designs are not widely used in medicine despite their potential to make improvement initiatives and health services research more efficient and effective. Quality managers, health system leaders, and directors of research institutes could increase productivity and gain important insights by promoting a broader use of factorial designs to study multiple interventions simultaneously and to learn from interactions.

Key words: Bradford Hill, clinical trial history, factorial, R. A. Fisher, randomized controlled trials, study designs

If there exists a pantheon for statisticians, it surely includes Sir Ronald Aylmer Fisher and Sir Austin Bradford Hill. Neither Fisher nor Hill cared for mathematical statistics without applications. Both wanted their methods to be of practical use and descended from Mount Olympus to help mere mortals in solving problems. While drawn to different domains, they maintained a collegial relationship until after Fisher's retirement.[1] But like the gods, they also had occasion to clash. Fisher is infamous for his scathing indictment[2] of Hill's landmark epidemiological research on health risks associated with the use of tobacco.[3] But less recognized is an earlier divergence between the 2 great scientists. It was not a conflict or formal disagreement; rather, Hill and Fisher held different mental models about practical application of state-of-the-art statistical methods.

One key difference in their perspectives concerned the use of sophisticated experimental designs such as factorial studies. A factorial design allows for simultaneous evaluation of 2 or more experimental factors (ie, treatments or interventions), as well as the interaction of those factors. Fisher recognized in 1926 that factorials would be an essential element of his contributions to experimental design:

> No aphorism is more frequently repeated in connection with field trials, than that we must ask Nature few questions, or, ideally, one question, at a time. The writer is convinced that this view is wholly mistaken. Nature, he suggests, will best respond to a logical and carefully thought out questionnaire; indeed, if we ask her a single question, she will often refuse to answer until some other topic has been discussed.[4(p511)]

Fisher encouraged multifactor designs in this early publication and throughout his career.[5] But "Fisher's message that factorial experiments can give you much more information . . . seems not to have been heeded" in medicine,[6(p940)] relative to the generally swift deployment of Fisher's multifactor studies in other industries and realms of academia.[7]

Path dependencies from early standards have generated a system that reinforces single-treatment trials. Fisher's message still seems little heeded in medicine, as factorial designs remain underutilized by clinical trial investigators, institutional funders, health services researchers, quality managers, and improvement leaders. This paradigm originated in part with a judgment call by Bradford Hill, when in the 1940s he played a key role in advancing the use of statistics in medicine. Hill believed that the medical professions would flinch from anything but the simplest experimental methods in the canon.[8,9] This judgment proved decisive when Hill went on to become his generation's foremost medical statistician and a founding father of the randomized controlled trial (RCT).[10] In a letter to Osler

Author Affiliations: *Department of Epidemiology and Biostatistics, School of Medicine, Case Western Reserve University, Cleveland, Ohio (Dr Neuhauser); Department of Information, Risk, and Operations Management, McCombs School of Business, The University of Texas at Austin (Dr Provost); Associates in Process Improvement, Austin, Texas (Mr Provost).*

Correspondence: *Shannon M. Provost, PhD, McCombs School of Business, 2110 Speedway, Stop B6500, Austin, TX 78712 (sprovost@utexas.edu).*

The authors have no conflicts of interest to declare.

Supplemental digital content is available for this article. Direct URL citation appears in the printed text and is provided in the HTML and PDF versions of this article on the journal's Web site (www.qmhcjournal.com).

Q Manage Health Care
Vol. 29, No. 2, pp. 109–122

DOI: 10.1097/QMH.0000000000000243

Peterson* on February 10, 1982, Hill described his reasoning:

> I became aware of Fisher's agricultural experiments as any statistician of that day must have been from his book *Statistical Methods for Research Workers*. But I was also aware that much of this work was elaborate with involved experimental designs and intricate analysis (of variance and co-variance, etc.). Any attempt to introduce these into clinical medicine would, in my opinion, have been fatal, and in this respect, **I deliberately turned my back on Fisher's methods**." [*emphasis added by authors*]

Hill helped to establish the single-factor RCT as a standard when he chose to promote the most straightforward study designs and to disregard Fisher as a source. The purpose of this article is to examine these events in context and to consider repercussions for evidence-based medicine. Other researchers have also contemplated why "Fisher's ideas about randomization and uncertainty had so little influence on medical understanding, then or now"[11(p933)] and considered how "Bradford Hill's understanding of medical susceptibility and ... concern for simplicity of design" were instrumental to his success in shaping the RCT as we know it.[12(p1220)] We contribute to these discussions and add new perspective on factorial designs. Hill's strategy in introducing experimental methods in early RCTs helped to create a methodological echo chamber around single-factor studies that still endures.

Our interest in the use of factorial designs was motivated in part by access to an unpublished recording of an interview between the lead author and Hill at the London Royal Society of Medicine on June 14, 1982. Also participating in the conversation was Philip D'Arcy Hart, MD.† This interview transcript and an unpublished 1982 letter from Hill offer new, first-person recollections of the early days of evidence-based medicine (see Supplemental Digital Content 1, published online, http://links.lww.com/QMH/A36). Relatively few of Hill's personal papers have survived for historians,[13] but some of his letters are available in his correspondents' collections such as the Fisher Archives in Adelaide.

HISTORY

We can better understand the surprising shortage of factorial studies in medical research by revisiting the careers of Fisher and Hill. Their interests, abilities, personalities, publications, and worldviews shaped the development of medical statistics, especially the RCT.

Sir Austin Bradford Hill FRS (1897-1991)

Bradford Hill was perhaps the world's preeminent medical statistician during the time of his remarkable career. His father, Leonard Hill FRS, was a prominent research physiologist and inaugural employee of the British Medical Research Council (MRC), the organization with which his son would eventually pioneer the RCT in the 1940s.[14] Bradford Hill had planned to study medicine, but the First World War interrupted his career plans and he contracted tuberculosis while serving in Greece. The Royal Naval Air Service dispatched Hill on a hospital ship and awarded him a full disability pension for life (an alarming prognosis).[15] Said Hill in a 1982 interview: "I got pneumothorax; a very early case—in 1917. ... And I got a lung abscess on top of it. ... It should have killed me, but it didn't."

Hill earned a London University correspondence degree in economics while convalescing.[16] Recovered by 1923, Hill pursued his interest in medicine with an MRC Institute position investigating industrial health problems.[17] A formative contact for Hill was Major Greenwood,[18] the leading medical statistician in early 20th-century Britain.[19] Greenwood had been mentored by Hill's father, and would return the favor by championing Hill at the MRC and guiding his career.[15] It was Greenwood who introduced Hill to statistics and encouraged him to attend Karl Pearson's lectures at University College London (UCL),[20] after which Hill "always acknowledged Pearson's influence on his own beliefs."[16(p483)]

When in 1927 Greenwood was appointed at the new London School of Hygiene and Tropical Medicine (LSHTM), he invited Hill to join his department.[8] Hill became a Reader in Epidemiology and Vital Statistics in 1933 and gained renown as an outstanding lecturer.[16] He is credited with introducing statistical methods to a generation of health care professionals.[17] In a 1982 interview, Hill described his work at the LSHTM:

> I got involved there and I was teaching ... largely what was entirely the first graduate medical students going into public health. And I got mixed up with a great many of MRC trials and several more MRC committees than anybody.

Hill would serve on 38 MRC committees between 1937 and 1970.[21] In 1945, Hill replaced Greenwood as LSHTM Chair of Medical Statistics and reassumed his work with the MRC Statistical Research Unit. In 1946,

*Osler Luther Peterson (1912-1988) was an American physician and researcher with a notable career at the Rockefeller Foundation, at the University of North Carolina's first population health department, and at Harvard Medical School.

†Philip Montagu D'Arcy Hart (1900-2006) was a British physician who dedicated his career to epidemiology and clinical research. He worked with the Medical Research Council (MRC) from 1937 to 1993, conducting in 1943 an early multisite trial of patulin treatment for the common cold and leading a series of tuberculosis vaccine trials involving 60 000 children. He was highly influential in introducing the RCT in medicine and deserves credit alongside Hill for propagating this innovation. Said Hill of his work with Hart at the MRC: "He argued from the medical point of view while I was arguing from the statistical."[8(p78)]

he began a program of research that many consider to be the first comprehensive RCT.[22] Hill's case control research that linked tobacco smoking with disease would further enhance his reputation.[23,24] Hill retired in 1961 but continued to consult and publish research. In 1965, he proposed a set of criteria for establishing causality[25] that remains an iconic epidemiological framework.

Sir Ronald Aylmer Fisher FRS (1890-1962)

R. A. Fisher was "the single most important figure in 20th century statistics."[26(p95)] Extreme myopia might have hindered Fisher's academic potential, but instead enhanced his mental reasoning skills (and kept him out of the First World War).[27] After distinguishing himself in mathematics and astronomy at Cambridge, he meandered through various teaching jobs at public schools and colleges.[28] Fisher's first publication in 1912 introduced the concept of maximum likelihood, an auspicious start to his extensive bibliography.[29] By 1914, he was corresponding with legendary mathematical statistician Karl Pearson and was able to solve within a week a problem—calculating the exact distribution of the correlation coefficient[30]—that had stymied Pearson and colleagues.[27]

In 1919, instead of pursuing Pearson's offer to become Chief Statistician at the Galton Lab, Fisher accepted a position at Rothamsted Experimental Station.[31] There, Fisher studied crop yields as impacted by combinations of fertilizers, chemicals, weather, and soil conditions.[32] Experimenting with agricultural plots led him to develop new statistical models and innovative methods such as learning from small data samples.[33] Fisher published his first book, *Statistical Methods for Research Workers*, in 1925.[34] It had 14 editions during his life and became a standard scientific reference. In 1935, Fisher published a foundational textbook, *The Design of Experiments*, and included a chapter introducing factorial study designs.[35] The entire field of experimental design and analysis can be traced back to this reference.

Fisher left Rothamsted in 1933 to replace Karl Pearson at UCL as Galton Chair of Statistics. He shared the position with Egon Pearson in a hostile departmental split that created a separate chair in Eugenics for Fisher,[36] one example of the extent to which Fisher's professional identity was shaped by his work in genetics and biology. Fisher's reputation grew in academic circles and his methods were increasingly adopted in diverse industries (although not, as we shall discuss, as extensively in medicine). Fisher frequently traveled to lecture and consult at institutions such as the University of Iowa, the Indian Statistical Institute in Calcutta, the US Department of Agriculture, the University of California at Berkeley, and the Commonwealth Scientific and Industrial Research Organization (CSIRO) in Adelaide, Australia.[37]

He was appointed Balfour Chair of Genetics in 1943 at the University of Cambridge, where he remained until his 1957 retirement.[38] Fisher returned as a research fellow to the CSIRO in 1959 and there concluded his career.[39] He was intellectually active until the week of his death in July 1962.[27]

Fisher and Hill: Shaping medical statistics

R. A. Fisher essentially invented the field of experimental design. So, it is conspicuous that he had so little bearing on the evolution of the RCT and puzzling that he faded into the background of the bourgeoning field of medical statistics during the peak of his career.[11] One explanation was that Fisher's own enthusiasm was reserved for other areas: namely, agriculture and eugenics and, above all, biology.[40] Fisher was considered a successor to Darwin based on his myriad contributions to biology and genetics.[41,42] It is interesting to note that biology (rather than statistics or mathematics) inspired Fisher's multifactor designs.[6] He modeled factorial studies on the simultaneous inheritance of Mendelian factors, also his origin for the name "factorial."[43]

Long-standing separation of (and sometimes tension between) medical practice and laboratory science[44] may have deterred clinicians who were interested in applying statistics to their work from turning to Fisher's references. Another barrier to awareness of Fisher's methods in medicine may have been his strong association with agriculture.[45] In a 1936 book review, *The British Medical Journal* praised Fisher's *The Design of Experiments* as "one of the most important contributions to scientific methodology of our generation," but also suggested that medical workers might be alienated by extensive agricultural examples and that "an easier introduction to these methods will find a larger public."[5(p365)]

Fisher's paradigm was quickly adopted by agricultural, biological, and industrial researchers but did not penetrate medical research (other than some limited animal-based experiments) until several decades later.[20] Instrumental in advancing Fisher's methods in medicine was Bradford Hill, although he failed to formally recognize Fisher as a source. Hill did not have a medical degree, but some believe he should have won a Nobel Prize in medicine for his work shepherding the RCT to become an international standard.[46] Because of Hill's role in popularizing medical statistics and his sway over the first generation of quantitatively fluent clinicians, his decision to ignore Fisher's methods was consequential.

Having established a reputation as an accessible lecturer and prolific writer, Hill was invited in 1937 to publish a series on medical statistics in *The Lancet*.[47] These articles were the basis of his textbook *Principles of Medical Statistics,*[48] which had 11 editions during Hill's lifetime. This book cemented his place in the medical statistics canon.[49] Hill's statistical content was at its core Fisherian, but his presentation style was notably different. By emphasizing practical examples and avoiding mathematical formulae, Hill "secured the attention of a largely innumerate medical profession."[21(p795)] Hill's writing and teaching were consistently praised for clarity, as though written by "a doctor interested in

statistics, not a statistician interested in medicine."[8(p103)] Said Hill in 1977: "My skill, if any, was ... offering the clinician something simple ... If one had started with something abstruse, the answer would have been 'Go to Hell'—and we would still be there."[50(p315)]

Hill did not reference Fisher in his *Lancet* articles or subsequent book,[49] but he was certainly aware of Fisher's work in the 1930s. Hill's colleague, J. Oscar Irwin, had worked at Rothamsted and was a known Fisher disciple.[1] Hill acknowledges Irwin's influence in his first edition, but he does not cite Fisher (other than to reference his χ^2 table). Hill added a chapter on the clinical trial to the 6th edition of *Principles of Medical Statistics* (1955) but again did not mention multifactor designs and failed to reference Fisher.[51] It seems that Hill viewed Fisher's methods as prohibitively complicated analyses that his less-sophisticated clinical audience might not tolerate,[20] as he indicated in a 1937 letter to Fisher.[6] Fisher responded on April 9, 1937:

> You are mistaken about the erudite character of the buyers who have made *Statistical Methods* a successful book. They are practical men, who want handy methods simply explained. I regard the legend that it is an advanced book as an injurious one, put about carelessly by some, and deliberately by others.[52]

And yet few would dispute that Fisher's books could be a "tough nut to crack."[13(p925)] Some complained that "a prerequisite for reading is ... a Master's degree in statistics."[53(p38)] Even William Gosset said of Fisher's writing: "when I come to 'evidently' I know that it means two hours hard work at least before I can see why."[54(p86)]

Fisher and Hill: Relationship and personalities

Bradford Hill was intentional in promoting basic experimental designs, but he did not disassociate his methods from those of Fisher because of a personal rivalry. In 1930, Fisher offered to nominate Hill as a Fellow of the Eugenics Society and hinted that he could seek a job at Rothamsted.[55] Hill wrote in 1933 to congratulate Fisher on his UCL appointment,[56] and reached out again in 1936 to request permission to include Fisher's table of χ^2 in his forthcoming book, *Principles of Medical Statistics*.[47] Fisher replied that he was "very glad to hear about your book ... certainly much needed."[52] In 1952, Hill personally informed Fisher that he would follow Hill as the next President of the Royal Statistical Society.[6] Fisher was known to stop by the LSHTM for friendly chats with Hill before meetings.[1] He also invited Hill to be his guest at the exclusive Royal Society Dining Club in 1954[57] and was said to be influential in securing Hill's election to the Royal Society.[58]

Fisher's relations with Hill were amicable, at least until the late 1950s.[59] But amity was not a trait generally associated with Fisher, who was widely known for toxic tendencies. Fisher had a complex character that was described as kind (particularly to junior scholars) but also caustic, and the latter impression tended to dominate.[6] Fisher's work was original to the extent that some pushback from established scholars was to be expected, but he was on occasion "quite appalling to people of more seniority against whom he had a grievance."[58(p947)] His precocity in mathematics and statistics notwithstanding, Fisher had been alienated early in his career from the inner circle of academic statisticians owing to an ongoing feud with Karl Pearson.[28] Pearson's influence was such that Fisher was relegated to the margins and compelled to publish in obscure journals in spite of his unequivocally innovative contributions and their relevance to the mathematical statistics mainstream.[60]

The feud with Karl Pearson was not entirely unfounded, but it was characteristic of Fisher's professional relationships. He also held a long-standing grudge with Polish statistician Jerzy Neyman. W. Edwards Deming described Fisher as a "'perfectly charming fellow' except when he was on the subject of Neyman."[61(p143)] Fisher "was not a great popularizer" and had a track record of hostility with subject matter experts (eg, at the Population Society and on a British Medical Association committee) who were seemingly reluctant to "engage with evidence."[11(p933)] Said physicist Raymond Birge: "he expects others to accept his discoveries without even questioning them."[61(p144)] Fisher's obstinance and lack of charisma made him singularly ill-suited to be a methodological champion in medicine.

While Fisher's "recalcitrant" personality was often an impediment,[37(p430)] Hill's character was central to his success. Hill—engaging, patient, and diplomatic—was an ideal statistical ambassador to the medical community.[12] He was a natural leader with a quiet humility that inspired junior colleagues.[16] Hill had long been interested in medicine and some have wondered if his near-death experience with tuberculosis was further inspiration for his career path.[58] Regardless, Hill made extensive and genuine efforts to understand the clinical perspective[62] and had the emotional intelligence and social skills to be effective in helping clinicians to manage uncertainty.[50] Hill commented in a 1982 interview (see Supplemental Digital Content 1, published online, http://links.lww.com/QMH/A36):

> John Crockton got me my honorary MD at Edinburgh, and I think he said, rightfully so, ... clinical trials ... John Linde did one in the 18th century ... but I sold them to a conservative profession.

Striking a balance between rigor and relevance, Hill was the right messenger and master of what his mentor Greenwood had called "statistical tact."[63(p155)] In a Royal Statistical Society lecture, Hill's colleague Peter Armitage recognized the challenge of introducing new ideas such as random allocation of treatments: "the noteworthy thing is not the time lag but the fact that randomization has become established at all in as difficult a subject as medicine."[64(p317)]

The dawn of the RCT

In 1946, the MRC established a Tuberculosis Research Unit (with Hill as a statistical adviser) to administer a trial for pulmonary tuberculosis patients.[65] The supervising MRC trials committee evidently gave Hill free rein to manage at the "front line," so Hill (along with colleagues Philip D'Arcy Hart and Marc Daniels) implemented "novel allocation by random sampling numbers."[66(p573)] Hill and colleagues, "engaged in a campaign to persuade as well as to explain," were busy convincing doctors to adopt the most basic concepts of experimental design, not to mention "Fisher's complex and subtle ideas."[67] In a 1982 letter, Hill elaborated on his rationale in promoting elementary methods:

> As a statistician one has to remember that the persons who carry out clinical trials are not usually statisticians or biostatisticians. They are clinicians who rightly like to know what they are doing, and why, and what the answer means. They have not invariably got at their fingertips the latest methods of statistical analysis, and an easily understood experiment and an easily understood analysis of results are more likely to ensure their cooperation and interest.

The streptomycin trial results were clear and required only elementary analysis, creating an ideal teaching case that helped to solidify the RCT model with participating physicians, most of whom may have been "uncomfortable with more complex approaches to interpreting experimental findings, especially when the results challenged established medical beliefs."[11(p935)] Said Hill in a 1982 interview:

> I wasn't going into complicated things like R.A. Fisher's official analysis . . . They weren't going to understand it. They wanted to know what the onset meant. They wanted to see what they were doing, of course, so it had to be very simple methods.

The trial results were published in 1948 and became an exceedingly influential model for future research.[68] Hill noted in 1963 that "many therapeutic trials in many branches of medicine have been founded upon this early essay."[69(p1043)] In 1982, Hill denied that Fisher's methods were a source for the streptomycin trial protocol:

> Well, nor did this come out of Fisher's teaching. I said no it didn't. I knew what Fisher was teaching with agricultural experiments. But it came out of what I was taught by Karl Pearson and Greenwood. We got these ideas in our heads all along before Fisher. Fisher was too elaborate anyway for medicine.

Because the streptomycin trial was widely acclaimed, so too became its single-factor design. But fixation on the streptomycin trial (at the expense of consideration of other pioneering studies) was somewhat arbitrary.‡ To be sure, the MRC investigators deserve special acknowledgment for design

Randomization in Medical Research

The use of randomization in experiments was another issue on which Fisher and Hill took different perspectives. Randomization was codified by Fisher in the 1920s as a basic tenet of experimental designs in agriculture.[34] Fisher showed how randomization could prevent bias in treatment selection and would also afford the use of probability in analysis.[35] Hill explained that in his *Lancet* articles he:

> deliberately left out the words "randomization" and "random sampling numbers" . . . trying to persuade doctors to come into controlled trials in the very simplest form and I might have scared them off . . . I thought it would be better to get doctors to walk first, before I tried to get them to run.[8(p77)]

Later, Hill used randomization to minimize bias in treatment assignments, but he never presented randomization as a key assumption associated with the statistical methods that he recommended, but rather as a practical measure to avoid selection bias in the trial.[65] In a 1982 interview, Hill elaborated on his practical motivations in the streptomycin trial:

> The entry was blind. . . . Mark Daniels said he didn't want the doctor to know whether the patient would get the treatment or not because then he might say, Oh, I won't put that patient in. That's when I did the sealed envelope. . . . I introduced the point that they shouldn't know what it is going to be.

One result of this framing is that many clinical researchers today do not conceptualize the assumption of randomization as fundamental to the validity of the statistical methods that they use. Fisher was clear that "the physical act of randomization is necessary for the validity of any test of significance."[35(p51)] Harry Marks' *The Progress of Experiment*[67] is an excellent reference on how the use of randomization has evolved in medical research.

‡As early as 1943, the MRC investigated patulin as a treatment for the common cold.[70] Some researchers believe that the patulin trial (in spite of its quasirandomization) merits wider recognition as one of the first modern double-blind trials with concurrent controls.[71] The illustrious streptomycin trial was neither double-blind nor placebo controlled,[72] two dimensions by which the patulin trial was methodologically superior.[66] The September 1946 streptomycin study was not even the first randomized MRC trial. Earlier in 1946, Hill used random sampling numbers in a prophylactic trial of a whooping cough vaccine, but its results were published after those of the streptomycin trial.[73]

intentionality, prescient ethical considerations, and an excellent report that clearly described steps taken to conceal from all participants pretrial knowledge of treatment allocation.[74] However, the streptomycin trial design and execution were not as singularly groundbreaking as would suggest its reputation.[12,75]

In 1946, a hepatitis trial supervised by the MRC used a factorial design to simultaneously study 2 dietary treatments that were allocated to alternate patients as a form of concurrent control.[76] This alternation approach presents a greater risk of selection bias than does full randomization,[74] a main reason that the fully randomized streptomycin trial received such attention. But why was the 1946 study not celebrated for its methodological innovation? If the hepatitis trial had been formally recognized for its complex (but more efficient) factorial design, other researchers might have followed suit with multifactor studies.

Hill's encouragement of single-factor designs was not limited to the MRC trials. The international institutionalization of the RCT owes much to his writing, consulting, and lecturing in the following decades.[77,78] Hill gave an invited lecture at Harvard Medical School in 1952, after which his talk was published by the *New England Journal of Medicine*.[79] Hill's work was also formative to leaders of early trials at the National Cancer Institute,[80] another example of his particular influence in introducing the RCT to American clinicians.[67] In 1982, Hill described his work in evangelizing RCTs:

> This is my life ... what I was involved in. ... So I did a great deal of advisory work. Again, in the early days, every man wasn't his own statistician. ... And I wanted to go along and give some advice.

Hill continued to lobby for simplicity in experimental design.[8] For example, during a 1953 lecture in the United States, Hill emphasized the basic designs used in early British trials and recommended that the National Institutes of Health simplify the complex design for its forthcoming multicenter trial.[81] In 1959, Hill was invited to chair a closed, 100-person Conference on Controlled Clinical Trials in Vienna, where he tried to codify best practices accrued in a decade of managing early MRC trials.[82] The published conference proceedings would become an RCT primer.[83]

Hill's students and colleagues took up his mantle and helped to further entrench the RCT method as a standard. Outstanding medical statisticians Peter Armitage and Donald Reid both considered Hill to be a key mentor. Richard Doll (Hill's student, friend, and MRC colleague) would become "synonymous with the conversion of modern clinical research to statistical models."[60(p187)] LSHTM student and MRC epidemiologist Archie Cochrane was known as a pioneer of evidence-based medicine and a prominent advocate of RCTs.[84] He inspired the international Cochrane Collaboration known for its formal organization of medical research findings.[85] As of 2019, the Cochrane Library database has more than 1.5 million records associated with hundreds of thousands of clinical trials.

DISCUSSION

Factorial study designs

Bradford Hill defined the RCT as "a carefully and ethically designed experiment with the aim of answering some precisely framed question."[86(p273)] This language suggests that the fundamental quest of a trial is pursuit of a *single answer*. Indeed, a typical RCT is designed to answer 1 question: Is Treatment X better than the usual care? If there are 200 patients in the study, 100 would be randomly assigned to the control group (receiving the usual care) with the other 100 getting Treatment X in the experimental group.

The factorial design starts with this plan but goes further. Factorials allow for estimation of the effects of multiple factors and their interactions (see the "Interactions in Factorial Designs" sidebar for an explanation of interaction effects). A card is created for each patient recording whether the patient is in the Treatment X experimental group or in the control group. All 200 cards are now combined and reshuffled into 2 new groups of 100 each. Out of these new groups, 1 group gets Treatment Y and the other group does not. We now have 4 groups:

1. 50 patients with Treatment X but not Treatment Y
2. 50 patients with Treatment Y but not Treatment X
3. 50 patients with both Treatment X and Treatment Y
4. 50 patients with neither Treatment X nor Treatment Y

By reshuffling, we can now answer 3 questions: Is Treatment X better? Is Treatment Y better? What about the combination of Treatment X and Treatment Y? We have potentially tripled the yield of information.

Why stop here? We might then make a note of each patient's group affiliation, combine the cards again, reshuffle, and then divide the patients into 2 new groups of size 100. Perhaps the new intervention is an exercise program (Z), in which 100 patients participate and 100 do not. Now we can answer 7 questions about 3 interventions, 3 pairwise interactions, and the 3-factor interaction. Is Treatment X, Y, or Z better? Is the combination of X+Y better than X+Z? Y+Z? X+Y+Z? We have potentially increased the yield of information by a factor of 7.

Why not keep going and divide the cards again? Why should we not study a dozen factors at a time? In principle this could be done. But a study's complexity increases along with its yield of information. Research questions in some environments warrant the inclusion of multiple factors. For example, a 2004 trial of postoperative nausea incorporated 6 factors.[87] This study was the first RCT in medicine that allowed for analysis of 3-factor interactions, providing reliable evidence to inform best practices in the use of anesthesia.[88] Another factorial study that rapidly changed medical practice standards was a 1988 trial of myocardial infarction treatments that revealed synergistic benefits

Interactions in Factorial Designs

In a cardiology trial for intracoronary infusions, Rentrop et al used a factorial design and identified a positive interaction between streptokinase and nitroglycerin. Patients assigned to this combined therapy had significant improvement in ejection fraction relative to those who received only 1 of the treatments, or neither (Figure 1 depicts a response plot that could have been used to display this important new finding).[91]

Fundamental to factorial designs is the distinction between main effects and interactions. A main effect describes the average impact of changing from 1 level of a factor to another. An *interaction* occurs when the effect of 1 factor in the study depends on the level or setting of another factor. Figure 2 illustrates the concept of an interaction with a hypothetical response plot. In graph A, each treatment influences the measure of interest, but these effects do not depend on the presence of the other factor. In graph B, each treatment effect depends on the presence of the other treatment. The effect of Treatment Y is 10 units when Treatment X is not present; whereas when X is present, the effect of Y is 40 units.

When an interaction is significant, the presentation of results of a study requires more elaboration. Estimates of main effects may not be useful, so the effect of each factor should be described with and without the presence of other factors with which it interacts. Response plots (along with 2-way tables) are a good method to display results of factorial studies.

Fisher's method to evaluate interactions in factorial studies is known as analysis of variance (ANOVA), an approach that tests for the presence of an interaction with the same statistical power as with testing for main effects.[35] Although ANOVA has been a popular method in clinical research, in the past, it was often not used with factorial designs. Of 83 factorial studies (compiled in 1991 from 3 prominent journals) reporting a significant interaction, only 24% used a complete and correct interpretation of the interaction using ANOVA.[92] This approach to reporting interaction effects may compromise valid interpretation of study results and could limit effective implementation of learning from medical studies.[92,93] We continue discussion of interaction effects in the section on obstacles to the use of factorial designs in medicine.

between oral aspirin and intravenous streptokinase.[89] In practice, 2- or 3-factor studies could be used in almost all clinical contexts.[90]

The consequences to society for stopping at one factor

Fixation on single-treatment trials has led to neglect of Fisher's factorial designs. Learning opportunities and cost savings have been lost along the way. As of June 2019, there were 309 645 studies registered on ClinicalTrials.gov (Figure 3 displays the number of studies in the database each year since 2000). A systematic review found that factorial RCTs accounted for only 4.6% of MEDLINE-listed trials from 1993 to 2003 (up from less than 1% during 1970-1980).[94] If the remaining 95.4% are single-factor studies, there have been more than 295 000 single-factor studies since the year 2000. Using Fisher's 3-factor design for each of these experiments could have answered 2 065 000 research questions (rather than merely the 295 000 that we have). If we assume that each single-factor study costs $1 million (not a stretch given that drug development trials may average $10 million each),[95] and if we value an answer at $100 000, we find that single-factor studies have forsaken $177 billion in answers. The reader is invited to speculate further. What might we have accomplished since the RCT emerged in the 1940s had factorial designs been used routinely?

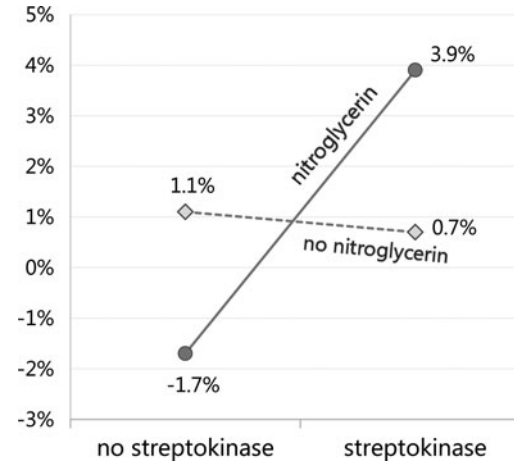

Figure 1. Response plot for change in ejection fraction.

Use of factorial designs in medicine

There are multiple reasons that factorial designs should be a strong consideration for an experimental study:

1. *Factorial designs allow for the evaluation of the interactions of each factor with all the other factors in the study.* As health care systems and disease management become more complex, it is increasingly important to understand interdependencies among interventions. Experimental treatments might have independent effects, or interact in a synergistic way, or interact in an antagonistic way (possibly neutralizing individual treatment effects). Factorial experiments are the only method that can estimate all possible interactions.[96] Ignoring potential interactions is a source of bias that could lead to the spread of treatments that do not make optimal use of health care resources.[97] Leveraging the understanding of interactions will increase precision of learning from studies and accelerate the rate of implementation of new knowledge.
2. *Factorial designs are more cost-efficient.* These designs are the most efficient study plan to learn about multiple factors because all data in a study are used to evaluate the significance of each of the factor and interaction effects. Studies that seek to detect small-to-moderate effects will require larger samples to observe the phenomena of

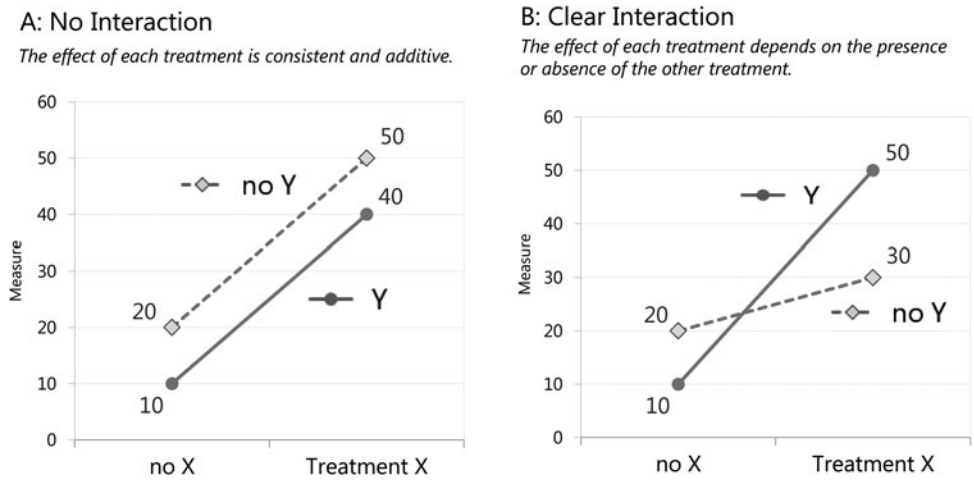

Figure 2. Response plots to illustrate the absence and presence of an interaction between treatments X and Y.

interest. Because large trials are time-consuming and expensive, investigators should exploit factorial designs offering valid answers to more than 1 question at a time.[98,99]

3. *Factorial designs are orthogonal.* Even with more than 2 interventions, the estimates of effects are independent of each other. There are no issues with confounding effects when using full factorial designs.
4. *Factorial designs encourage a comprehensive approach to learning about complex systems.* Thinking about multiple interventions gives researchers a broader system view of their focal problems. If they have a primary factor of interest, it will be evaluated over a wide range of the other factors in the study. Results will be more robust when applied in new environments. Multifactor experiments are well-suited to shed light on the underlying change mechanisms associated with the experimental research questions.[100]

While extensive studies with many factors present practical limitations, fractional factorial designs are an attractive option to keep the experiment at a reasonable size.[7,101] Factorial designs may also be used sequentially to make sense of complex problems. Collins et al[102] developed a methodological framework involving a series of randomized experiments for screening, refining, and confirming intervention components; the key method for screening and refining is factorial analysis.

Factorial studies (mostly 2-factor) have appeared infrequently in the medical literature over the last 75 years.[103] The Table lists illustrative examples of multifactor designs used in research and improvement studies. We curated this list to acquaint the reader with examples of factorial studies addressing a variety of clinical topics and published in different journals between 1946 and 2019.

We are not the first researchers to lobby for factorials in quality improvement studies[116] and in medical research. In 1952, an insightful text by Donald Mainland clearly outlined factorial design and analysis. Having studied with Fisher during the 1930s,[117] he made a clear endorsement to clinicians seeking quantitative guidance:

> A medical student may not see any immediate prospect of using these designs; but there are three reasons why he should know about them:
>
> 1. No one who is unaware of these methods know what the term "modern statistics" means.

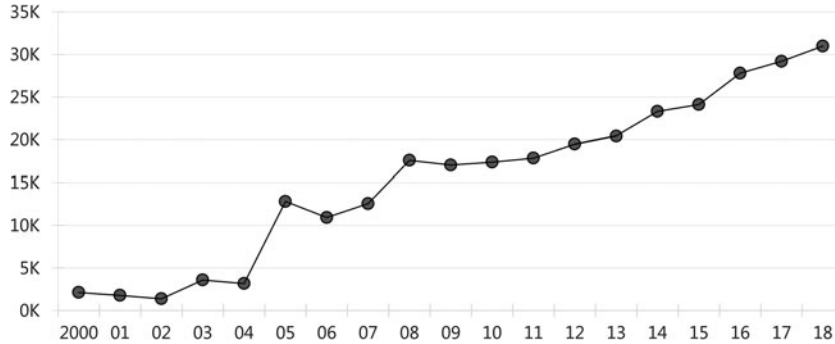

Figure 3. Studies registered at ClinicalTrials.gov during 2000-2018.

Table. Some Noteworthy Factorial Study Designs in the Medical Literature

Year	Journal	Title	Design
1946	*The Lancet*	Diet in the Treatment of Infective Hepatitis: Therapeutic Trial of Cysteine and Variation of Fat-Content[76]	2^2 factorial
1955	*Journal of Clinical Investigation*	The Treatment of Acute Infectious Hepatitis. Controlled Studies of The Effects of Diet, Rest, and Physical Reconditioning on the Acute Course of the Disease and on the Incidence of Relapses and Residual Abnormalities[104]	2^2 factorial and 2^3 factorial
1960	*The British Medical Journal*	Stilboestrol, Phenobarbitone, and Diet in Chronic Duodenal Ulcer: A Factorial Therapeutic Trial[105]	2^3 factorial
1976	*The Lancet*	Effects of Timolol And Hydrochlorothiazide on Blood-Pressure and Plasma Renin Activity[106]	2^2 factorial
1988	*The Lancet*	Randomised Trial of Intravenous Streptokinase, Oral Aspirin, Both, or Neither Among 17187 Cases of Suspected Acute Myocardial Infarction: ISIS-2[89]	2^2 factorial
1989	*Journal of the American College of Cardiology*	Late Thrombolytic Therapy Preserves Left Ventricular Function in Patients With Collateralized Total Coronary Occlusion: Primary End Point Findings of the Second Mount Sinai-New York University Reperfusion Trial[107]	2^2 factorial
1998	*Journal of American Medical Association*	Lumbar Supports and Education for the Prevention of Low Back Pain in Industry, an RCT[91]	2^2 factorial
2001	*The Archives of Ophthalmology*	Diabetes and Postoperative Endophthalmitis in the Endophthalmitis Vitrectomy Study[108]	2^2 factorial
2002	*British Medical Journal*	Randomised Factorial Trial of Falls Prevention Among Older People Living in Their Own Homes[109]	2^3 factorial
2004	*New England Journal of Medicine*	A Factorial Trial of Six Interventions for the Prevention of Postoperative Nausea and Vomiting (IMPACT)[87]	2^6 factorial
2007	*Journal of American Medical Association*	Effects of Citalopram and Interpersonal Psychotherapy on Depression in Patients With Coronary Artery Disease[110]	2^2 factorial
2008	*American Journal of Public Health*	Screening Experiments and the Use of Fractional Factorial Designs in Behavioral Intervention Research[111]	Fractional factorial (multiphase)
2012	*Pediatrics*	Improving Notification of Sexually Transmitted Infections: A Quality Improvement Project and Planned Experiment[112]	2^2 factorial (2 replications)
2016	*Journal of American Medical Association*	Effect of Behavioral Interventions on Inappropriate Antibiotic Prescribing Among Primary Care Practices, a RCT[113]	2^3 factorial
2016	*Pediatrics*	SLUG Bug: Quality Improvement with Orchestrated Testing Leads to NICU CLABSI Reduction[114]	2^{4-1} factorial
2019	*BMJ Quality & Safety*	Effect of Two Behavioural "Nudging" Interventions on Management Decisions for Low Back Pain: A Randomised Vignette-Based Study in General Practitioners[115]	2^2 factorial

2. The methods will be increasingly met in reports on medical research.
3. Unless one knows what an efficiently designed experiment will do, one cannot realize how defective are the old-fashioned methods still in use.[118(p194)]

Thomas Chalmers, RCT scholar and meta-analytic innovator,[119] led an early factorial RCT of hepatitis treatments for soldiers in the Korean War.[104] His 1955 trial report was commended[120] and remains among exemplary factorial studies in the literature (Table). Chalmers was a strong proponent of factorial designs and noted the lack of consideration of factorial trials in medicine and biostatistics:

> It seems to me that the cardiovascular and the cancer clinical trial people are drastically underutilizing a very useful technique which could save a lot of time, effort and money. . . . there are not many people who use this technique. I find the major problem lies with the biostatisticians who suffer from some kind of extreme bias against it. I have tried to sell the technique to a number of people on a number of occasions . . . it is the only way available to detect interaction, pharmacologically a most important phenomenon.[90(p286)]

Sir Richard Peto pointed out that multifactor trials are "more valuable scientifically . . . one of the few substantial improvements in clinical trial design which can be implemented with little or no extra difficulty or cost."[121(p34)] Peto—enthusiastic about factorials for their efficiency, capacity to reveal interactions, and potential to encourage collaboration among researchers—asked in 1978: "Surely such designs should be commoner than they now are?"[121(p35)]

Although not himself a factorial advocate, there is reason to believe that Bradford Hill recognized the benefits of more complex study designs. At his suggestion, some of Hill's LSHTM colleagues explored extensions of the basic RCT, including a factorial design to be used when "treatments are likely to act by different mechanisms."[10(p1524)] Doll presented on multifactor study designs at the landmark RCT conference in Vienna chaired by Hill in 1959.[83] Doll wrote in 2005 that factorials had "become established as a valuable technique that has enabled conclusions to be drawn . . . much more quickly and more cheaply."[98(p480)] But the "establishment" of factorial designs seems to have been a premature conclusion in 2005. Of the 1135 RCTs indexed in PubMed during December 2000 and December 2006, only 14 studies (1.2%) used a factorial design.[122] As we demonstrate, researchers are still are not using factorial designs at anywhere near the scale and scope for which they are appropriate.

Some obstacles to the use of factorials

When *The British Medical Journal* reviewed Fisher's *The Design of Experiments*[5] in 1936, they emphasized an illustrative quote on the use of multifactor designs:

> If single factors are chosen for investigation, it is not because we anticipate that the laws of nature can be expressed with any particular simplicity in terms of these variables, but because they are variables which can be controlled or measured with comparative ease. If the investigator, in these circumstances, confines his attention to any single factor, we may infer either that he is the unfortunate victim of a doctrinaire theory as to how experimentation should proceed, or that the time, material, or equipment at his disposal are too limited to allow him to give attention to more than one narrow aspect of his problem.[35(p97)]

Here, Fisher suggests that an investigator's decision to study a single factor may be rooted in a fundamental misconception about the nature of scientific inquiry—that we learn effectively by studying only 1 variable at a time. This fallacy has perpetuated the single-factor RCT as a gold standard, as have other forces. The use of suboptimal methods in factorial trials and too-frequent misinterpretation of results have perpetuated a myth about diminished statistical power to detect interaction effects. Furthermore, many of the published factorial studies that do evaluate interactions fail to do so properly,[93] leading many researchers to deliberately steer away from factorial trials because the presence of an interaction is widely portrayed as a methodological menace.

We have already elaborated on a key barrier to the use of factorials: Hill did not reference Fisher's methods in his most-cited writings.[1,67] Fisher had little direct bearing on the RCT in medicine and "even less on the way physicians were taught to understand statistics."[11(p935)] Building on Hill's proposed methods, many early investigators overlooked multifactor designs and failed to invoke Fisher as a reference. Another hurdle is the scant treatment of advanced experimental designs in medical education references. Many textbooks[123-125] offer minimal coverage of experimental methods more advanced than the basic 1-factor study (with some notable exceptions[118,126]). Too frequently, medical students are not exposed to opportunities offered by multifactor designs.

Some references introduce factorial study designs but provide little guidance on analyzing the results of a factorial experiment; others adequately summarize factorial analysis but miss the point on some aspect of Fisher's method. This is particularly glaring with respect to interaction effects (see the "Interactions in Factorial Designs" sidebar). The importance of interactions in understanding complex systems is not fully appreciated in the medical literature. Too often researchers indicate only that no significant interaction was present, neglecting to report numerical results such that readers could form their own assessment about the presence of an interaction.[127] Even Richard Doll, an early advocate of factorials, omits any discussion of interactions in his 2005 summary paper on multifactor trials.[98] Results from a 2003 systematic review suggest that 1 in 3 published articles on factorial trials do not include analysis of interactions.[94] A common mistake is to report only within-subgroup probability values for treatment effects rather than conduct statistical tests of interaction.[128] Many articles that report statistically significant interactions fail to correctly interpret those effects[129] or use the wrong underlying model,[130] perpetuating another myth about factorial designs and limited statistical power.

Numerous factorial studies that appear in the literature fail to use the methodologically indicated ANOVA by Fisher.[131] Instead, a common suboptimal practice for factorial analysis has historically been to use a χ^2 test of the average of the main effects and to test for the interaction using only half of the data.[127] In fact, a researcher using ANOVA tests the significance of an interaction with the same power as the test of a main effect. It is only after a meaningful interaction is revealed that estimates of interaction effects are made with less precision.[35] In the absence of a significant interaction, researchers may proceed with marginal analyses of main treatment effects.[132] In other words, if the effect of 1 factor does not vary as a function of the other, then the average effect of each factor may be calculated with half the sample that would be required for separate studies of each factor.[104]

These common misconceptions about statistical power and interactions have bolstered yet another unfortunate phenomenon limiting the use of factorials. Many researchers consider factorial designs only when they anticipate no interaction between experimental treatments.[96] Again, readily available RCT references substantiate these rumors. The Cochrane Collaboration Handbook[133] offers this discussion on interactions:

> In most factorial trials the intention is to achieve "two trials for the price of one," and the assumption is made that the effects of the different active interventions are independent, that is, there is no interaction (synergy). Occasionally a trial may be carried out specifically to investigate whether there is an interaction between two treatments.

A unique advantage of factorials is the capacity to observe interactions, so there is no need to assume independence of the interventions when designing a factorial research study. A 2003 systematic review in the *Journal of the American Medical Association* promoted this inaccurate view in suggesting that "factorial trials are ideal when the 2 treatments act independently" and by concluding that factorially designed "investigations are appropriately choosing to test only those interventions that do not have potential for substantive interaction."[94(p2545)] Other recent publications echo these claims, extending the view of interactions as something to avoid.[93]

Research proposals are often more focused on a specific intervention rather than the outcome of interest. It is easier to provide the science and theory for testing a single-factor hypothesis than discuss a hypothesis for multiple factors. Therefore, researchers are naturally encouraged to focus their designs on isolating the impact of this specific intervention as opposed to understanding how the intervention works in a larger system of variables affecting an outcome. Quality improvement studies are usually focused on changing an outcome; thus, it is more natural to think about multiple factors that could combine to potentially change the outcome. In the context of both research and improvement, it is simpler to test an intervention when no interaction is present. But this is not an argument for ignoring interactions or failing to characterize them. If a researcher suspects that interventions may interact, it should be imperative to study the interventions together in a factorial design. "Errors associated with . . . interaction effects constitute a threat to the statistical conclusion validity of medical research"[92(p1571)] and the willful inattention to interactions is detrimental to good science.

Armitage called in 1979 for more factorials but may have actually discouraged their use by adding "although they give rise to difficulties of interpretation when factors interact."[134(p266)] More recently, Armitage coauthored an exemplary medical textbook[126] that offers thorough coverage of factorial studies, as do recent articles.[100,132] But even these modern references remain exceptional rather than typical with respect to discussion of factorial design and analysis. In other industries, popular references on experimental design have consistently invoked Fisher's methods.[7,135-140]

CONCLUSION

It is time to reconsider the role of factorial designs in both medical research and improvement studies. Bradford Hill's choice to put forward the most accessible study designs was plausible in the 1940s. Indeed, medical professionals would likely have been spooked by complex methods, hindering the advancement of rigorous and systematic research. Medical statistics was a nascent discipline at the time; health care organizations did not routinely staff analysts, nor were any physicians expected to be familiar with quantitative methods. Computational power would have been an issue, along with management of large data sets. But statistical analyses now used regularly in RCTs (such as hierarchical models and logistic regression) are computationally more complex than anything Fisher proposed.

Hill's reasoning to promote single-factor studies does not hold up today. Statistical applications in medicine are thriving, medical institutions are teeming with new digital data sources, and computational power has increased exponentially. Health care offers diverse career opportunities for data scientists, and clinicians themselves are expected to learn different statistical methods. It follows that factorial designs should be used much more frequently. Factorial studies are efficient, requiring a smaller sample than would separate trials of each factor or a 3-arm trial with both factors and a control group. The factorial is the only study design that can identify a potential interaction between factors. That factorial experimenters must endure loss of statistical power in detecting the presence of interactions is a pernicious myth. Clinical research and improvement studies could be more effective—*exploiting interactions*—and efficient—*testing multiple interventions*—if we routinely considered the use of factorial experiments.

Researchers, clinical scientists, editors, grant agencies, academic leaders, textbook authors, and quality improvement managers should all examine their role in facilitating the use of factorial designs in research and improvement studies. Medical students need to be exposed early in their careers to references that accurately describe the design and analysis of factorial studies (such as that by Armitage and colleagues).[126]

Organizations funding research grants and sponsoring quality improvement initiatives have leverage to encourage the use of factorials. Institutional websites should encourage the consideration of multifactor designs in proposals and offer guidelines for describing theories and hypotheses for factorial designs. An abundance of publications based on 1-factor studies has preserved the notion that a trial should investigate only 1 treatment. There is no medical subject heading term specified for factorial trials in the MEDLINE database, making reports from these studies harder to find and reference.[94] Those who lead quality improvement in health care delivery systems often work under budgetary constraints and face tough choices in prioritizing improvement projects and choosing where to dedicate limited resources for quality initiatives. Methodological guidelines for factorials in the health care improvement literature (as well as the incorporation of experimental design into improvement training programs) would expand the portfolios of those managing improvement

at different levels of scale and present attractive study designs to enhance learning productivity.

The greater use of factorial designs will lead to efficiencies, but insights that come from studying interactions may be even more important to improving health and health care systems.[141] With many patients taking multiple drug regimens and increasingly complex combined interventions, the clinical practice landscape is teeming with interactions. Yet the evidence for interactions is relatively anecdotal and often spread by word-of-mouth or side effect registries. We could rapidly improve at learning from interactions with greater use of multifactor studies.

Bradford Hill's substantial contributions to medical statistics are irrefutable. Under his influence, RCTs became the regular instrument for rigorous learning in medicine. Perhaps it is now time for R. A. Fisher's methods to drive innovation on this front. The vast majority of trials in 7 decades of RCT history have investigated a single intervention. What more we might have learned from widespread use of factorial trials is history, but we can rethink the role of factorial designs for the future.

REFERENCES

1. Armitage P. Fisher, Bradford Hill, and randomization. *Int J Epidemiol*. 2003;32(6):925-928.
2. Fisher RA. Dangers of cigarette-smoking. *Br Med J*. 1957; 2(5039):297.
3. Doll R, Hill AB. Lung cancer and other causes of death in relation to smoking. *Br Med J*. 1956;2(5001):1071.
4. Fisher RA. The arrangement of field experiments. *J Min Agric Great Britain*. 1926(33):503-513.
5. Review Editors. The planning of experiments. *Br Med J*. 1936; 1:365-365.
6. Bodmer W. RA Fisher, statistician and geneticist extraordinary: a personal view. *Int J Epidemiol*. 2003;32(6):938-942.
7. Moen RD, Nolan TW, Provost LP. *Quality Improvement Through Planned Experimentation*. New York, NY: McGraw-Hill; 1999.
8. Silverman WA, Chalmers I. Sir Austin Bradford Hill: an appreciation. *Control Clin Trials*. 1992;13(2):100-105.
9. Hill AB. Memories of the British streptomycin trial in tuberculosis: the first randomized clinical trial. *Control Clin Trials*. 1990; 11(2):77-79.
10. Doll R. Sir Austin Bradford Hill and the progress of medical science. *Br Med J*. 1992;305(6868):1521.
11. Marks HM. Rigorous uncertainty: why RA Fisher is important. *Int J Epidemiol*. 2003;32(6):932-937.
12. Doll R. Controlled trials: the 1948 watershed. *Br Med J*. 1998; 317(7167):1217-1220.
13. Chalmers I. Fisher and Bradford Hill: theory and pragmatism? *Int J Epidemiol*. 2003;32(6):922-924.
14. Hill AB, Hill B. The life of Sir Leonard Erskine Hill FRS (1866-1952). *Proc R Soc Med*. 1968;61(3):307-316.
15. Hill I. Austin Bradford Hill—ancestry and early life. *Stat Med*. 1982;1(4):297-300.
16. Armitage P. Obituary: Sir Austin Bradford Hill, 1897-1991. *J R Stat Soc Ser A Stat Soc*. 1991;154(3):482-484.
17. Himsworth SH. Bradford Hill and statistics in medicine. *Stat Med*. 1982;1(4):301-303.
18. Doll R. Sir Austin Bradford Hill: a personal view of his contribution to epidemiology. *J R Stat Soc Ser A Stat Soc*. 1995;158(1):155-163.
19. Farewell V, Johnson T. Major Greenwood (1880-1949): a biographical and bibliographical study. *Stat Med*. 2016;35(5):645-670.
20. Armitage P. Before and after Bradford Hill: some trends in medical statistics. *J R Stat Soc Ser A Stat Soc*. 1995;158(1):143-153.
21. Doll R. Sir Austin Bradford Hill, 1897-1991. *Stat Med*. 1993; 12(8):795-808.
22. Medical Research Council. Streptomycin treatment of pulmonary tuberculosis. *Br Med J*. 1948;2:169-782.
23. Doll R, Hill AB. Smoking and carcinoma of the lung. *Br Med J*. 1950;2(4682):739.
24. Doll R, Hill AB. The mortality of doctors in relation to their smoking habits. *Br Med J*. 1954;1(4877):1451.
25. Hill AB. The environment and disease: association or causation? *Proc R Soc Med*. 1965;58(5):295-300.
26. Efron B. RA Fisher in the 21st century. *Statist Sci*. 1998;13(2):95-114.
27. Yates F, Mather K. Ronald Aylmer Fisher, 1890-1962. *Biogr Mem Fellows R Soc*. 1963;9:91-129.
28. Box JF, Edwards A. Fisher, Ronald Aylmer. In: Armitage P, Colton T, eds. *Encyclopedia of Biostatistics*. New York, NY: John Wiley & Sons; 2005.
29. Fisher RA. On an absolute criterion for fitting frequency curves. *Mess Math*. 1912;41:155-156.
30. Fisher RA. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*. 1915;10(4):507-521.
31. Hald A. *A History of Mathematical Statistics From 1750 to 1930*. New York, NY: Wiley; 1998.
32. Box JF. RA Fisher and the design of experiments, 1922–1926. *Am Stat*. 1980;34(1):1-7.
33. Box JF. Gosset, Fisher, and the t distribution. *Am Stat*. 1981; 35(2):61-66.
34. Fisher RA. *Statistical Methods for Research Workers*. Edinburgh, Scotland: Oliver and Boyd; 1925.
35. Fisher RA. *The Design of Experiments*. Edinburgh, Scotland: Oliver and Boyd; 1935.
36. Lenhard J. Models and statistical inference: the controversy between Fisher and Neyman Pearson. *Br J Philos Sci*. 2006; 57(1):69-91.
37. Box JF. *R. A. Fisher, the Life of a Scientist*. New York, NY: Wiley; 1978.
38. Fernandes D. From Three Fishers: Statistician, Geneticist and Person to Only One Fisher: The Scientist. *J Biom Biostat*. 2016;7:282.
39. Ludbrook J. R.A. Fisher's Life and Death in Australia, 1959-1962. *Am Stat*. 2005;59(2):164-165.
40. Fisher RA. *The Genetical Theory of Natural Selection*. Oxford, England: Oxford University Press; 1999.
41. Edwards A. Mathematizing Darwin. *Behav Ecol Sociobiol*. 2011; 65(3):421-430.
42. Box JF. Commentary: on RA Fisher's Bateson lecture on statistical methods in genetics. *Int J Epidemiol*. 2010;3(2):335-339.
43. Fisher R. Statistical methods in genetics: the Bateson Lecture, 1951. *Heredity*. 1952;6:1-12.
44. Swales J. The troublesome search for evidence: three cultures in need of integration. *J R Soc Med*. 2000;93(8):402-407.
45. Fisher RA. Design of experiments. *Br Med J*. 1936;1(3923):554-554.
46. Lock S. The randomised controlled trial—a British invention. In: Lawrence GM, ed. Technologies of Modern Medicine. London, England: Science Museum; 1994:81-87.
47. Farewell V, Johnson A. The origins of Austin Bradford Hill's classic textbook of medical statistics. *J R Soc Med*. 2012;105(11):483-489.
48. Hill AB. *Principles of Medical Statistics*. London, England: The Lancet; 1937.
49. Farewell V, Johnson T. Woods and Russell, Hill, and the emergence of medical statistics. *Stat Med*. 2010;29(14):1459-1476.
50. Schoolman HM. The clinician and the statistician. *Stat Med*. 1982;1(4):311-316.
51. Hill AB. *Principles of Medical Statistics*. 6th ed. London, England: The Lancet; 1955.
52. Fisher RA. Correspondence with AB Hill, April 9, 1937. Barr Smith Library, The University of Adelaide. http://hdl.handle.net/2440/67745. Accessed July 29, 2019.
53. Rao CR. RA Fisher: The founder of modern statistics. *Statist Sci*. 1992;7(1):34-48.

54. Bodmer WF. Genetic sequences. *Proc R Soc Lond B Biol Sci*. 1990;241(1301):85-92.
55. Fisher RA. Correspondence with AB Hill, December 2, 1930. Barr Smith Library, The University of Adelaide. http://hdl.handle.net/2440/67745. Accessed July 29, 2019.
56. Fisher RA. Correspondence with AB Hill, June 14, 1937. Barr Smith Library, The University of Adelaide. http://hdl.handle.net/2440/67745. Accessed July 29, 2019.
57. Fisher RA. Correspondence with AB Hill, January 4, 1954. Barr Smith Library, The University of Adelaide. http://hdl.handle.net/2440/67745. Accessed July 29, 2019.
58. Dickersin K, Armitage P, Djulbegovic B, et al. Fisher and Bradford Hill: a discussion. *Int J Epidemiol*. 2003;32(6):945-948.
59. Stolley PD. When genius errs: RA Fisher and the lung cancer controversy. *Am J Epidemiol*. 1991;133(5):416-425.
60. Salsburg D. *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*. New York, NY: Henry Holt and Company; 2001.
61. Reid C. *Neyman—From Life*. New York, NY: Springer; 1998.
62. Hill AB. Alfred Watson Memorial Lecture: the statistician in medicine. *J Inst Actuar*. 1962;88(2):178-191.
63. Greenwood M. Is the statistical method of any value in medical research. *Lancet*. 1924;2:153-158.
64. Armitage P. Statistical methods in clinical and preventive medicine. *J R Stat Soc Ser A Stat Soc*. 1963;126(2):316-317.
65. Chalmers I. Statistical theory was not the reason that randomisation was used in the British Medical Research Council's clinical trial of streptomycin for pulmonary tuberculosis. In: Jorland G, Weisz G, Opinel A, eds. *Body Counts: Medical Quantification in Historical and Sociological Perspectives*. Montreal, Canada: Fondation Me'rieux by McGill-Queen's University Press; 2005:309-334.
66. Hart PDA. A change in scientific approach: from alternation to randomised allocation in clinical trials in the 1940s. *BMJ*. 1999;319(7209):572-573.
67. Marks HM. *The Progress of Experiment: Science and Therapeutic Reform in the United States, 1900-1990*. Cambridge, England: Cambridge University Press; 2000.
68. Daniels M. Scientific appraisement of new drugs in tuberculosis. *Am Rev Tuberc*. 1950;61(5):751-756.
69. Hill AB. Medical ethics and controlled trials. *Br Med J*. 1963;1(5337):1043.
70. Medical Research Council Patulin Clinical Trials Committee. Clinical trials of patulin in the common cold. *Lancet*. 1944;2:373-375.
71. Chalmers I, Clarke I, M. Commentary: the 1944 patulin trial: the first properly controlled multicentre trial conducted under the aegis of the British Medical Research Council. *Int J Epidemiol*. 2004;33(2):253-260.
72. Doll SR. Clinical trials: retrospect and prospect. *Stat Med*. 1982;1(4):337-344.
73. Medical Research Council Whooping-cough Immunization Committee. The prevention of whooping-cough by vaccination. *BMJ*. 1951;1:1463-1471.
74. Chalmers I. Why transition from alternation to randomisation in clinical trials was made. *BMJ*. 1999;319(7221):1372.
75. Yoshioka A. Use of randomisation in the Medical Research Council's clinical trial of streptomycin in pulmonary tuberculosis in the 1940s. *BMJ*. 1998;317(7167):1220-1223.
76. Wilson C, Pollock M, Harris A. Diet in the treatment of infective hepatitis: therapeutic trial of cysteine and variation of fat-content. *Lancet*. 1946;247(6407):881-883.
77. Chalmers I. UK Medical Research Council and multicentre clinical trials: from a damning report to international recognition. *J R Soc Med*. 2013;106(12):498-509.
78. Crofton J. The MRC randomized trial of streptomycin and its legacy: a view from the clinical front line. *J R Soc Med*. 2006;99(10):531-534.
79. Hill AB. The clinical trial. *N Engl J Med*. 1952;247(4):113-119.
80. Gehan EA, Schneiderman MA. Historical and methodological developments in clinical trials at the National Cancer Institute. *Stat Med*. 1990;9(8):871-880.
81. Hill A. *The Philosophy of the Clinical Trial: National Institute of Health Annual Lectures–1953*. Washington, DC: Public Health Service; 1953.
82. Bird SM. The 1959 meeting in Vienna on controlled clinical trials—a methodological landmark. *J R Soc Med*. 2015;108(9):372-375.
83. Doll R. The Concurrent Assessment of Several Treatments. In: Hill AB, ed. *Controlled Clinical Trials*. Oxford, England: Blackwell Scientific Publications; 1960:87-93.
84. Cochrane AL, Blythe M. *One Man's Medicine: An Autobiography of Professor Archie Cochrane*. London, England: British Medical Journal; 1989.
85. Chalmers I. The Cochrane collaboration: preparing, maintaining, and disseminating systematic reviews of the effects of health care. *Ann N Y Acad Sci*. 1993;703(1):156-165.
86. Hill AB. *Principles of Medical Statistics*. 9th ed. New York, NY: Oxford University Press; 1971.
87. Apfel CC, Korttila K, Abdalla M, et al. A factorial trial of six interventions for the prevention of postoperative nausea and vomiting. *N Engl J Med*. 2004;350(24):2441-2451.
88. Korttila K, Apfel CC. Factorial design provides evidence to guide practice of anaesthesia. *Anaesthesiol Scand*. 2005;49(7):927-929.
89. Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17,187 cases of suspected acute myocardial infarction: ISIS-2. ISIS-2 (Second International Study of Infarct Survival) Collaborative Group. *Lancet*. 1988;2(8607):349-360.
90. Chalmers TC. A potpourri of RCT topics. *Control Clin Trials*. 1982;3(3):285-298.
91. van Poppel MN, Koes BW, van der Ploeg T, Smid T, Bouter LM. Lumbar supports and education for the prevention of low back pain in industry: a randomized controlled trial. *JAMA*. 1998;279(22):1789-1794.
92. Ottenbacher KJ. Interpretation of interaction in factorial analysis of variance design. *Stat Med*. 1991;10(10):1565-1571.
93. Pocock SJ, Hughes MD, Lee RJ. Statistical problems in the reporting of clinical trials. *N Engl J Med*. 1987;317(7):426-432.
94. McAlister FA, Straus SE, Sackett DL, Altman DG. Analysis and reporting of factorial trials: a systematic review. *JAMA*. 2003;289(19):2545-2553.
95. DiMasi JA, Hansen RW, Grabowski HG. The price of innovation: new estimates of drug development costs. *J Health Econ*. 2003;22(2):151-185.
96. Montgomery AA, Peters TJ, Little P. Design, analysis and presentation of factorial randomised controlled trials. *BMC Med Res Methodol*. 2003;3(1):26.
97. Dakin H, Gray A. Economic evaluation of factorial randomised controlled trials: challenges, methods and recommendations. *Stat Med*. 2017;36(18):2814-2830.
98. Doll R. Controlled trials testing two or more treatments simultaneously. *J R Soc Med*. 2005;98(10):479-480.
99. Stampfer MJ, Buring JE, Willett W, Rosner B, Eberlein K, Hennekens CH. The 2 × 2 factorial design: its application to a randomized trial of aspirin and US physicians. *Stat Med*. 1985;4(2):111-116.
100. Baker TB, Smith SS, Bolt DM, et al. Implementing clinical research using factorial designs: a primer. *Behav Ther*. 2017;48(4):567-580.
101. Chakraborty B, Collins LM, Strecher VJ, Murphy SA. Developing multicomponent interventions using fractional factorial designs. *Stat Med*. 2009;28(21):2687-2708.
102. Collins LM, Murphy SA, Strecher V. The multiphase optimization strategy (MOST) and the sequential multiple assignment randomized trial (SMART): new methods for more potent eHealth interventions. *Am J Prev Med*. 2007;32(5 suppl):S112-S118.
103. Chan AW, Altman DG. Epidemiology and reporting of randomised trials published in PubMed journals. *Lancet*. 2005;365(9465):1159-1162.
104. Chalmers TC, Eckhardt RD, Reynolds WE, et al. The treatment of acute infectious hepatitis. *J Clin Invest*. 1955;34(7):1163-1235.
105. Truelove, S. Stilboestrol, phenobarbitone, and diet in chronic duodenal ulcer. *Br Med J*. 1960;2(5198):559.
106. Chalmers J, Tiller D, Horvath J, Bune A. Effects of timolol and hydrochlorothiazide on blood-pressure and plasma renin activity. *Lancet*. 1976;2(7981):328-331.
107. Rentrop KP, Feit F, Sherman W, et al. Late thrombolytic therapy preserves left ventricular function in patients with collateralized

total coronary occlusion: primary end point findings of the Second Mount Sinai-New York University Reperfusion Trial. *J Am Coll Cardiol*. 1989;14(1):58-64.
108. Doft BH, Wisniewski SR, Kelsey SF, Fitzgerald SG. Diabetes and postoperative endophthalmitis in the endophthalmitis vitrectomy study. *Arch Ophthalmol*. 2001;119(5):650-656.
109. Day L, Fildes B, Gordon I, Fitzharris M, Flamer H, Lord S. Randomised factorial trial of falls prevention among older people living in their own homes. *BMJ*. 2002;325(7356):128.
110. Lespérance F, Frasure-Smith N, Koszycki D, et al. Effects of citalopram and interpersonal psychotherapy on depression in patients with coronary artery disease: the Canadian Cardiac Randomized Evaluation of Antidepressant and Psychotherapy Efficacy (CREATE) trial. *JAMA*. 2007;297(4):367-379.
111. Nair V, Strecher V, Fagerlin A, et al. Screening experiments and the use of fractional factorial designs in behavioral intervention research. *Am J Public Health*. 2008;98(8):1354-1359.
112. Huppert JS, Reed JL, Munafo JK, et al. Improving notification of sexually transmitted infections: a quality improvement project and planned experiment. *Pediatrics*. 2012;130(2):e415-e422.
113. Meeker D, Linder JA, Fox CR, et al. Effect of behavioral interventions on inappropriate antibiotic prescribing among primary care practices: a randomized clinical trial. *JAMA*. 2016;315(6):562-570.
114. Piazza AJ, Brozanski B, Provost L, et al. SLUG bug: quality improvement with orchestrated testing leads to NICU CLABSI reduction. *Pediatrics*. 2016;137(1):e20143642.
115. Soon J, Traeger AC, Elshaug AG, et al. Effect of two behavioural "nudging" interventions on management decisions for low back pain: a randomised vignette-based study in general practitioners. *BMJ Qual Saf*. 2019;28(7):547-555.
116. Speroff T, O'Connor GT. Study designs for PDSA quality improvement research. *Qual Manag Health Care*. 2004;13(1):17-32.
117. Altman D. Donald Mainland: anatomist, educator, thinker, medical statistician, trialist, rheumatologist. JLL Bulletin: Commentaries on the history of treatment evaluation. https://www.jameslindlibrary.org/articles/donald-mainland-anatomist-educator-thinker-medical-statistician-trialist-rheumatologist/ Published 2017.
118. Mainland D. *Elementary Medical Statistics: The Principles of Quantitative Medicine*. 6th ed. London & Philadelphia: WB Saunders Co; 1952.
119. Dickersin K, Chalmers K, F. Thomas C Chalmers (1917–1995): a pioneer of randomised clinical trials and systematic reviews. *J R Soc Med*. 2015;108(6):237-241.
120. Sackett D. A 1955 clinical trial report that changed my career. *J R Soc Med*. 2010;103(6):254.
121. Peto R. Clinical trial methodology. *Biomedicine*. 1978;28:24-36.
122. Hopewell S, Dutton S, Yu LM, Chan AW, Altman DG. The quality of reports of randomised trials in 2000 and 2006: comparative study of articles indexed in PubMed. *BMJ*. 2010;340:c723.
123. Meinert CL, Tonascia S. *Clinical Trials: Design, Conduct and Analysis*. Vol 39. 8th ed. New York, NY: Oxford University Press; 1986.
124. Bland M. *An Introduction to Medical Statistics*. 2nd ed. Oxford, England: Oxford University Press; 1995.
125. Fletcher R, Fletcher S, Wagner E. *Clinical Epidemiology: The Essentials*. 2n ed. Baltimore, MD: Williams and Wilkins; 1988.
126. Armitage P, Berry G, Matthews JNS. *Statistical Methods in Medical Research*. Hoboken, NJ: John Wiley & Sons; 2008.
127. Lubsen J, Pocock S. Factorial trials in cardiology: pros and cons. *Eur Heart J*. 1994;15:585-588.
128. Rothwell PM. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *Lancet*. 2005;365(9454):176-186.
129. Marascuilo LA, Levin JR. Appropriate post hoc comparisons for interaction and nested hypotheses in analysis of variance designs: the elimination of type IV errors. *Am Educ Res J*. 1970;7(3):397-421.
130. Levin JR, Marascuilo LA. Type IV errors and interactions. *Psychol Bull*. 1972;78(5):368-374.
131. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet*. 2000;355(9209):1064-1069.
132. Sedgwick P. What is a factorial study design? *BMJ*. 2014;349:g5455.
133. Green S, Higgins J, Alderson P, Clarke M, Mulrow C, Oxman A. *Cochrane Handbook for Systematic Reviews of Interventions*. West Sussex, England: John Wiley & Sons Ltd; 2008.
134. Armitage P. The design of clinical trials. *Aust J Stat*. 1979;21(3):266-281.
135. Box GE, Hunter WG, Hunter JS. *Statistics for Experimenters*. New York, NY: John Wiley and Sons; 1978.
136. Snedecor GW, Cochran WG. *Statistical Methods*. 7th ed. Ames, IA: Iowa State University Press; 1980.
137. Winer BJ, Brown DR, Michels KM. *Statistical Principles in Experimental Design*. New York, NY: McGraw-Hill; 1971.
138. Campbell DT, Stanley JL. *Experimental and Quasi-Experimental Designs for Research*. Chicago, IL: Rand McNally; 1966.
139. Cox DR. *Planning of Experiments*. New York, NY: Wiley; 1958.
140. Cochran W, Cox G. *Experimental Designs*. 2nd ed. New York, NY: Wiley; 1957.
141. Berlin JA. What are factorial experiments and why can they be helpful? *JAMA Netw Open*. 2019;2(9):e1911917.