The Royal Society of Medicine

Journal of the Royal Society of Medicine; 2021, Vol. 114(5) 264–270
DOI: 10.1177/01410768211003790

Richard Cabot, pair-matched random allocation, and the attempt to compare like with like in the social sciences and medicine. Part 2: the context of medicine and public health

Scott H Podolsky[1], Brandon C Welsh[2] and Steven N Zane[3]
[1]Department of Global Health and Social Medicine, Harvard Medical School, Boston, MA 02115, USA
[2]School of Criminology and Criminal Justice, Northeastern University, Boston, MA 02115, USA
[3]College of Criminology and Criminal Justice, Florida State University, Tallahassee, FL 32304, USA
Corresponding author: Scott H Podolsky. Email: scott_podolsky@hms.harvard.edu

Introduction

As described in Part 1, Richard Cabot's 1937 Cambridge-Somerville Youth Study on the impact of a social intervention of 'directed friendship' on youth delinquency represented the first trial in the social or behavioural sciences to use alternate or random allocation after matching study participants into pairs. Cabot himself, as an esteemed physician, straddled the worlds of both the social sciences and medicine. Part 1 described the study and placed it within the historical context of the social sciences. We next turn to the historical context of medicine and public health.

The advent of matching

Within medicine, from the turn of the 20th century onward and despite the abundance of uncontrolled studies of interventions, some researchers began to consider how to compare like with like in clinical experiments. Many such researchers, often investigating the prophylaxis or treatment of infectious diseases, employed 'alternate allocation' studies, in which patient A would receive one intervention, patient B an alternative or nothing, and so forth (other versions of this entailed alternating the treatment plan every other day, or, in an anticipation of cluster randomisation, alternating wards of a hospital to treatment versus control groups). Dozens of such studies were conducted during the first half of the 20th century.[1] By the 1930s and 1940s, concerns over researchers believing that they could 'improve on' allocation schedules based on alternation or random allocation, for example by preferentially steering the sickest patients to the novel treatment, led Austin Bradford Hill to conceal allocation schedules from those entering patients as participants in controlled trials.[2,3] Concealed allocation schedules in the 1948 Medical Research Council study of streptomycin for pulmonary tuberculosis contributed to its subsequent iconic status in the history of treatment evaluation.[4]

But there were other methods offered for ensuring fair medical treatment comparisons. One important technique was that of 'matching', or attempting to ensure, a priori rather than solely in post-hoc analysis, equivalent representation of seemingly relevant characteristics among treated and untreated groups. To some extent, such notions extend to James Lind's own assertion that the cases of the sailors in his experiment comparing different treatments for scurvy 'were as similar as I could have them'.[5] By the early 20th century, some trialists demanded increasing attention to ensuring such matched characteristics.

One articulation of intentional matching as a method per se in the evaluation of therapeutics appears in a 1912 paper by Harry Lee Barnes on the treatment of tuberculosis with tuberculin, although this was a retrospective analysis.[6] Superintendent of the State Sanatorium in Rhode Island, Barnes conducted a retrospective analysis of 150 patients treated at the sanatorium between 1907 and 1912. As he stated, comparisons

> should be drawn between two classes of patients, those who take the treatment and those who do not, and these parallels should be made from cases that are as similar in prognosis as possible. For this study, an attempt was made to match each one of the 150 patients taking tuberculin against another patient of the same classification, according to the National Association, and also anatomically

> according to Turban, and likewise to match only cases having similar records of bacilli in the sputum, temperature, pulse, respiration, general condition, weight, race and year of discharge.

Finding no evidence of benefit of tuberculin, Barnes nevertheless considered that his methodology itself represented an advance:

> While not perfect it should be much superior to slipshod methods of stating results of treatment and if widely adopted it would help to weed out more rapidly worthless methods of treatment in pulmonary tuberculosis. If applied to mooted questions like the 'value of climate,' it would eventually solve them, as the fruitless war of theories and opinions would eventually be displaced by evidence.

Nevertheless, Barnes presciently acknowledged: 'Drawbacks to the use of this method are the abundance of material required and the amount of labour necessary to carry it out'.

We see emphasis on the need for equivalence around key factors in *prospective* studies in such prominent locations as Major Greenwood's and Udny Yule's World War I-era paper on 'The Statistics of Anti-Typhoid and Anti-Cholera Inoculation, and the Interpretation of Such Statistics in General'[7] and in the subsequent American Public Health Association 'Working Program against Influenza'. As John Eyler has pointed out, in the wake of the mass of uncontrolled (or poorly controlled, even to contemporary judges) influenza vaccination studies during the influenza pandemic, the 'Working Program' authors stated as one of their key characteristics of a valid study comparing vaccinated to unvaccinated individuals that 'the relative susceptibilities of the two groups should be equal, as measured by age and sex distribution', as well as exposure history.[8,9] There is no mention of alternate allocation (let alone random allocation) in the 'Working Program'. Indeed, it did not otherwise specify how such groups were to be rendered equivalent with respect to such factors.

More formal attention to a priori matching using key characteristics appeared in several prominent prospective studies throughout the 1920s. In Harriette Chick et al.'s investigation of the influence of diet and sunlight on rickets in institutionalised children in postwar Vienna,

> the children on admission were placed in two groups upon Diets I and II, care being taken that the infants in each group should be as similar as possible in age, general condition, and development, and that they should remain under identical conditions of general management and hygiene during their stay in the hospital.[10,11]

To further ensure equivalence with respect to environmental exposure, 'the children in the two dietetic groups occupied adjoining cots in the wards, so that differences in the degree of illumination and exposure to fresh air were minimised as much as possible'.

In Elmer McCollum's controlled study of supplementary milk in 84 institutionalised children in Baltimore divided into treatment and control groups, 'every effort was made … so that any child in one group was comparable in age, size and condition to a child in the other group'.[12,13] Likewise, in Harold Corry Mann's study of milk supplementation among institutionalised children outside London, the division of children into active or control groups took account of age, as well as a combined rating score of height and weight.[14,13]

In a study of vaccines for preventing common colds at the University of Manchester, researchers took 144 volunteers 'and divided them into two equal groups by sorting the cards [filled out by the volunteers] first according to the sex of the volunteers, and then according to the dates on which the last cold was recorded'.[15] As they continued, emphasising the characteristics they were seemingly able to control versus those they were not able to: 'Thus the two groups were approximately alike with regard to sex-distribution and with regard to the period which had elapsed since the last cold, in all other respects the distribution was random'. In none of the four studies was there any mention of how participants were allocated to active versus control groups.

Matching and stratification in alternate allocation studies

By contrast, among those focusing on alternate allocation as a means of ensuring the comparison of like with like, matching prior to allocation could be described as an impractical luxury, especially when researchers felt that with strict alternation and a large enough sample size, important characteristics would distribute sufficiently evenly among active versus control groups. Patients with pneumonia at Harlem Hospital were given polyvalent antiserum (treating multiple pneumococcal serotypes), and 'because of the importance of treating patients at the earliest moment it was impracticable to alternate [the patients by pneumococcal serotype], since often at least twelve hours would have been lost before this

was determined'.[16] William Park, Jesse Bullowa and Milton Rosenblüth, conducting the study, 'believed that with a sufficiently large series the distribution of case by type would be equalised between the treated and the untreated group', and indeed this proved to be the case. Similarly, when the British Medical Research Council began its own study of anti-pneumococcal antiserum a few years later, they clearly enunciated exclusion criteria (e.g. no patients with advanced heart disease, no patients under the age of 20 or over the age of 60) to avoid confounding factors, but their plan

> still left altogether unregulated the chance scatter of distribution of patients with severe or mild pneumonia into either the serum or control groups, and also of those admitted for treatment early or relatively late in the progress of the disease.[17]

As with the American pneumonia researchers, it 'was thought better not to attempt a deliberate sorting of cases in respect of mildness or severity, but to trust that differences due to chance scatter would become almost negligible in a fairly large number of cases'. However, analysis of the Medical Research Council trial noted that in some of the participating sites more 'severe' cases had ended up in the treated groups rather than in the control groups, evidence that set the stage for future discussions of the limitations of unconcealed allocation schedules.[18]

To some extent, such divisions between the use of a priori stratified studies and alternate allocation may be considered to represent the practical differences between planned, slowly enrolling studies of chronic conditions or preventive measures, and interventions in acute illnesses like pneumonia. But certain researchers did take pains to carefully stratify patients into subgroups for comparison *before* alternate allocation took place. In the early 1920s, Nicholas Kopeloff and George Kirby, at the New York State Psychiatric Institute on Ward's Island, investigated the impact of the elimination of focal infections (dental, tonsillar or cervical) on psychiatric illness.[19,20] As they noted, 'because of the difficulties of interpretation inherent in an investigation of this nature, it seemed desirable to reduce the study as nearly as possible to the terms of an experiment'. They chose alternate allocation as their primary means for ensuring equivalence between treated and untreated patients, but also noted that 'an attempt was made to place in the two different groups, patients comparable as to sex, age, duration of psychosis, diagnosis, prognosis, and infective conditions'. It is unclear how exactly they attempted to operationalise this methodological foreshadowing of 'minimization'[21] or reconcile it with alternate allocation.

A decade later, Massachusetts General Hospital's Donald King, studying the inhalation of carbon dioxide to prevent postoperative pulmonary complications, was far more explicit in describing his attempt to stratify patients prior to alternate allocation. He began by noting that

> since the sex of the patient and the type of abdominal operation play so important a part, the patients were divided according to sex and then grouped according to the type of abdominal operation. Every other patient, in the subgroups of each sex, was treated.[22]

He continued:

> This alternation gave, for instance, a group of men who had had operations on the stomach and who had had hyperventilation induced, to compare with an equal number of men who had had operations on the stomach but who had not had hyperventilation induced. ... Thus, statistics were available for male and female cases, treated and untreated, in the different groups of abdominal operations and hernia repair.

Random allocation within matched pairs

Such general tensions between matching and alternate allocation would be paralleled among those who first broached the mixed application of matching and random allocation within medicine and public health, bringing us still closer to Cabot's study. By the 1920s, Ronald Fisher had advocated random allocation among agriculture plots in 'The Arrangement of Field Experiments'.[23] Ian Hacking has noted that, in contrast, 'a majority of traditionalists believed that "matched" or "balanced" arrangements were less subject to error, more instructive, and in general entitled one to draw firmer instances',[24] and that William Sealy Gosset (a traditionalist who published under the pseudonym 'Student', and originator of 'Student's t-test') eventually favoured 'balanced randomization' as a happy compromise.

This played out in a fascinating way in 1930 and 1931 with respect to a 'nutritional experiment on a very large scale' that followed upon the milk studies described above.[13,25] In Lanarkshire, Scotland, 20,000 students from 67 schools were studied in the spring of 1930 to assess the effects of milk supplementation on growth. In any given school, for the

most part, 'the teachers selected the two classes of pupils, those getting milk and those acting as ''controls'', in two different ways. In certain cases they selected them by ballot and in others on an alphabetical system'.[26] However, 'in any particular school where there was any group to which these methods had given an undue proportion of well-fed or ill-nourished children, others were substituted in order to attain a more level selection'. In other words, a rough form of 'matching' was added to the process to ensure the comparison of like with like. The study seemed to favour the inclusion of milk; but most important to our inquiry, by 1931, it had led Gosset to produce a methodological deconstruction of the study.

For Gosset, foreshadowing the concerns of those who revealed well-intentioned cheating with unconcealed allocation schedules, 'unconscious selection' (later in the paper referred to as 'unconscious bias'), seemingly manifested in the attempt at matching, could lead to the production of unequal comparison groups (as seemed to have been the case in the Lanarkshire study). Especially focusing on a subquestion of the study concerning the relative utility of raw versus pasteurised milk, Gosset noted that the studied students 'were not random samples from the same population; they were selected samples from populations which may have been different, ... [and] I would be very chary of drawing any conclusions from these small biased differences'. As he gently lamented, 'this experiment, in spite of all the good work which was put into it, just lacked the essential condition of randomness which would have enabled us to prove the fact'. Instead, Gosset proposed that if the experiment were to be repeated 'on the same spectacular scale', then:

> The 'controls' and 'feeders' should be chosen by the teachers in pairs of the same age group and sex, and as similar in height, weight and especially physical condition (i.e. well or ill nourished) as possible, and divided into 'controls' and 'feeders' by tossing a coin for each pair.

In a subsequent section of the paper, concerning the comparison between raw versus pasteurised milk, Gosset noted that among 20,000 children, there should be, on average, about 50 pairs of identical twins and that 'the error of the comparison between them may be relied upon to be so small that 50 pairs of these would give more reliable results than the 20,000 with which we have been dealing'. Again, he proposed a plan whereby the researchers would '''[f]eed'' one of each pair on raw and the other on pasteurised milk, deciding in each case by the toss of a coin which is to take raw milk'. Richard Cabot's constructed 'diagnostic twins', discussed in Part 1, had been foreshadowed by such literal twins.

That same year witnessed the publication by J Burns Amberson et al. of 'A Clinical Trial of Sanocrysin in Pulmonary Tuberculosis'.[27] Twenty-four patients 'free from serious complications' participated in the study:

> On the basis of clinical, X-ray and laboratory findings the 24 patients were divided into two approximately comparable groups of 12 each. The cases were individually matched, one with another, in making this division. Obviously, the matching could not be precise, but it was as close as possible, each patient having previously been studied independently by two of us.

Finally, 'by a flip of the coin, one group became identified as group I (sanocrysin-treated) and the other as group II (control)'.

Amberson et al.'s study did not uncover any beneficial effects of sanocrysin; indeed the drug was shown to have nasty side effects. Joseph Gabriel has demonstrated the origins of the trial at the intersection of mutual public health service and pharmaceutical industry (Parke Davis) interest in an objective assessment of the drug, with the trial entailing blinding of patients to prevent a 'psychic influence' on healing.[28] More germane to our line of inquiry, the origins of the single coin toss to determine the allocation of the two groups of patients are less apparent from the archival record. George McCoy, who had played a large role in the American Public Health Association vaccine protocols that emphasised matching (as mentioned above), also supported this therapeutic trial through his role as the director of the national Hygienic Laboratory. While the expressed need for a controlled study (even in discussions of the animal studies that preceded the human study) is evident throughout the record, and while the 'plan' for the trial initially called for 100 treatment and 50 control patients, there is no formal mention in the plan of either matching (beyond the intent to choose groups of patients 'on the basis of pulmonary lesions that are as nearly as possible comparable as regards extent and character of disease') or random allocation.[29] Clearly, however, by the late 1920s and early 1930s, certain trialists were extending beyond would-be matched controls to the addition of random allocation as an additional mechanism to ensure fair comparisons between treatment and control groups.

Conclusions

Despite our extensive searching within the Cabot and Sheldon and Eleanor Glueck papers at Harvard (the Glueck's had a close relationship with Cabot and were influential in Cabot's development of the Cambridge-Somerville Youth Study), it is unclear whether Cabot was aware of Kopeloff and Kirby's trial, King's study, Gosset's dissection of the Lanarkshire study, Amberson et al.'s tuberculosis trial or Austin Bradford Hill's discussion of the 'Principles of Medical Statistics' in *The Lancet* in 1937. On the one hand, Cabot conducted research on tuberculosis in his early medical career and wrote about the disease and its treatment in his medical textbooks, including later editions published after 1931.[30–32] On the other hand, we have not found reference to any of these studies, either in his medical publications or in his personal notes or correspondence.

In tracing the history of treatment evaluation and the conduct of fair comparisons, it would seem that there is more of a direct line from the advent of alternate allocation, through concerns over their improper implementation, to the advent of randomised clinical trials, than there is from Amberson, Cabot or even Gosset to Bradford Hill.[2] In this reading, the mixed matching plus randomisation proposals and studies of the 1920s and 1930s seem to be a relative dead end, albeit one reflecting increasing concern to provide objective assessment of novel interventions in the interwar years. These developments ensured that like would be compared with like, unmeasured variables would be unbiasedly distributed among comparison groups, and that, by concealing the allocation schedule, the allocation system itself could not be cheated.

The combination of matching with random allocation in prospective clinical trials would continue to be deployed in both the social sciences and medicine throughout the 20th and into the 21st centuries, followed by evolving debate over its advantages and limitations.[33–36] The design itself serves as a cornerstone of the evolving articulation of stratification, matching, randomisation, and similar innovations for ensuring fair comparisons are made in trials. Key to this history has been Richard Cabot's Cambridge-Somerville Youth Study, the first large-scale matched-randomised trial and one of the earliest randomised trials of a social intervention.[37]

Prior research pointed to the study's design as representing a natural carry-over from Cabot's background in clinical and research medicine, as well as the design appealing to him on the grounds that it would be even more rigorous than alternation or simple random allocation.[38] We found additional support for the influence of the latter, with Cabot viewing matching as insufficient on its own to achieve equivalence between treatment and control groups. Equally compelling was Cabot's added concern about 'achieving adequate experimental controls' in evaluating an intervention that focused on social behaviour. The implication is that the social world compared to the physical world was less known to experimentalists. This is to take nothing away from Cabot's staunch advocacy for social workers and his repeated call for them to evaluate their interventions using rigorous methods. Famously, in his presidential address to the National Conference of Social Work, he made clear his desire for an age of rigorous evaluation in social work, 'the much-to-be-desired epoch when we shall control our results by comparison with a parallel series of cases in which we did nothing'.[39]

More broadly, our research has situated both Cabot and his study in the midst of the social sciences and medicine and public health as they have wrestled with the uses of a priori stratification, matching, alternate allocation, and random allocation, and attempted to compare like with like in the 20th and 21st centuries. Matching – whether as an independent form of ensuring seemingly unbiased comparisons or as an a priori component of alternate allocation or random allocation – has so far received insufficient attention (as have related notions of stratification and exclusion criteria). Our research is an attempt to address this historical gap. Additionally, we have tried to place the history of the social sciences, medicine and public health in direct conversation with one another. The boundaries among such disciplines are indeed indistinct and dynamic. For example, the 1916 study cited in Part 1 on the role of air quality in education was overseen by the New York State Commission on Ventilation, with noted public health pioneer Charles-Edward Amory Winslow among the Commission's listed members (Thorndike and Ruger 1916).[40] 'Cross-ventilation' among the social sciences, medicine and public health themselves has persisted to this day. Richard Cabot likely served as only the most prominent of individuals who straddled – or at least engaged with – multiple disciplines. We hope historians will follow suit and that these articles will stimulate further attention in this direction.

Declarations

Competing Interests: None declared.

Funding: None declared.

Ethics approval: Not applicable.

Guarantor: SHP.

Contributorship: Each of the authors participated in all parts of study development and writing of the manuscript.

Acknowledgements: We are especially grateful to the reference staff at Harvard University Archives, Harvard Law School Library's Historical and Special Collections and the Center for the History of Medicine at the Countway Library. We thank Joseph Gabriel for generously making available more than 500 images of pages from the sanocrysin files held at the National Archives. Finally, we wish to thank Iain Chalmers for suggesting the central questions that guided our research, as well as for his sage advice and insightful comments throughout the development of this articles.

Provenance: Invited article from the James Lind Library.

References

1. Chalmers I, Dukan E, Podolsky SH and Smith GD. The advent of fair treatment allocation schedules in clinical trials during the 19th and early 20th centuries. *J R Soc Med* 2012; 105: 221–227. See https://www.jameslindlibrary.org/articles/the-advent-of-fair-treatment-allocation-schedules-in-clinical-trials-during-the-19th-and-early-20th-centuries/ (last checked 25 February 2021).
2. 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, Opinel A, Weisz G, eds. *Body Counts: Medical Quantification in Historical and Sociological Perspectives*. Montreal: McGill-Queens University Press, 2005: 309–334.
3. Bothwell LE and Podolsky SH. The emergence of the randomized, controlled trial. *N Eng J Med* 2016; 375: 501–504.
4. Medical Research Council. Streptomycin treatment of pulmonary tuberculosis. *Br Med J* 1948; 2: 769–782.
5. Lind J. *A Treatise of the Scurvy. In Three Parts. Containing an Inquiry into the Nature, Causes, and Cure, of that Disease. Together with a Critical and Chronological View of what has been Published on the Subject*. Edinburgh: Sands, Murray, and Cochran for A. Kincaid and A. Donaldson, 1753.
6. Barnes HL. Report of 150 cases of pulmonary tuberculosis treated with tuberculin. *J Am Med Assoc* 1912; 59: 332–333.
7. Greenwood M and Yule GU. The statistics of anti-typhoid and anti-cholera inoculation, and the interpretation of such statistics in general. Proc R Soc Med *(Sect Epidemiol State Med)* 1914–1915; 8: 113–190.
8. Evans WA, Armstrong DB, Davis WH, Kopp EW and Woodward WC. A working program against influenza. *Am J Public Health*, 1919; 9: 1–13.
9. Eyler J. The fog of research: influenza vaccine trials during the 1918-19 pandemic. *J Hist Med Allied Sci* 2009; 64: 401–428.
10. Chick H, Dalyell EJ, Hume EM, Mackay HMM, Henderson Smith H and Wimberger H. *Study of rickets in Vienna, 1919-1922*. Report to the Accessory Food Factors Committee appointed jointly by the Medical Research Council and the Lister Institute. Medical Research Council Special Report Series No. 77. London: HMSO, 1923: 19–94.
11. Chick H. Study of rickets in Vienna, 1919–1922. *Med Hist* 1976; 20: 41–51.
12. McCollum EV. The nutritional value of milk. In: Rogers LA, Lenoir RD, eds. *World's Dairy Congress, Washington, D.C., 2–10 October 1923*. Washington: US Government Printing Office, 1924: 421–427.
13. Pollock J. Two controlled trials of supplementary feeding of British school children in the 1920s. *J R Soc Med* 2006; 99: 323–327. See https://www.jameslindlibrary.org/articles/two-controlled-trials-of-supplementary-feeding-of-british-school-children-in-the-1920s/ (last checked 25 February 2021).
14. Corry Mann HC. *Diets for Boys During the School Age*. Medical Research Council Special Report Series No. 105. London: HMSO, 1926.
15. Ferguson FR, Davey AFC and Topley WWC. The value of mixed vaccines in the prevention of the common cold. *J Hyg (Lond)* 1927; 26: 98–109.
16. Park WH, Bullowa JGM and Rosenblüth MB. The treatment of lobar pneumonia with refined specific antibacterial serum. *J Am Med Assoc* 1928; 91: 1503–1508.
17. Medical Research Council. The serum treatment of lobar pneumonia. *Br Med J* 1934; 1; 241–245.
18. Chalmers I. UK Medical Research Council and multicentre clinical trials: from a damning report to international recognition. JLL Bulletin: Commentaries on the history of treatment evaluation, 2013. See https://www.jameslindlibrary.org/articles/uk-medical-research-council-and-multicentre-clinical-trials-from-a-damning-report-to-international-recognition/ (last checked 25 February 2021).
19. Kopeloff N and Kirby GH. Focal infection and mental disease. *Am J Psychiatry* 1923; 3: 149–197.
20. Wessely S. Surgery for the treatment of psychiatric illness: The need to test untested theories. JLL Bulletin: Commentaries on the history of treatment evaluation, 2009. See https://www.jameslindlibrary.org/articles/surgery-for-the-treatment-of-psychiatric-illness-the-need-to-test-untested-theories/ (last checked 25 February 2021).
21. Pocock SJ and Simon R. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics* 1975; 31: 103–115.
22. King DS. Postoperative pulmonary complications. II. Carbon dioxide as a preventive in a controlled series. *J Am Med Assoc* 1933; 100: 21–26.
23. Fisher RA. The arrangement of field experiments. *J Min Agric Gr Br* 1926; 33: 503–513.
24. Hacking I. Telepathy: origins of randomization in experimental design. *Isis* 1988; 79: 427–451.
25. Gosset WS. The Lanarkshire milk experiment. *Biometrika* 1931; 23: 398–406.
26. Leighton G and McKinlay P. *Milk Consumption and the Growth of School Children*. Edinburgh/London: Department of Health for Scotland/HMSO, 1930.

27. Amberson JB, McMahon BT and Pinner M. A clinical trial of sanocrysin in pulmonary tuberculosis. *Am Rev Tuberc* 1931; 24: 401–435.
28. Gabriel JM. The testing of sanocrysin: science, profit, and innovation in clinical trial design, 1926–31. *J Hist Med Allied Sci* 2014; 69: 604–632.
29. Plan for Clinical Test of Sanocrysin, November 1926, RG 443, General Records of the NIH, 1930-1948, box 21, ''Sanocrysin Clinical Tests.''
30. Cabot RC. Research Society Reports, conference on tuberculosis of the lungs. *War Med* 1919; 2: 978–979.
31. Cabot RC. *A Layman's Handbook of Medicine: With Special Reference to Social Workers*, 2nd ed. Boston: Houghton Mifflin, 1937.
32. Cabot RC and Adams FD. *Physical Diagnosis*, 12th ed. Baltimore: Williams and Wilkins, 1938.
33. Billewicz WZ. Matched samples in medical investigations. *Brit J Prev Soc Med* 1964; 18: 167–173.
34. Bland JM and Altman DG. Matching. *BMJ* 1994; 309: 1128.
35. Farrington DP and Welsh BC. A half century of randomized experiments on crime and justice. *Crime Justice* 2006; 34: 55–132.
36. Ariel B and Farrington DP. Randomized block designs. In: Piquero AR, Weisburd D, eds, *Handbook of Quantitative Criminology*. New York: Springer, 2010: 437–454.
37. Forsetlund L, Chalmers I and Bjørndal A. When was random allocation first used to generate comparison groups in experiments to assess the effect of social interventions? *Econ Innov New Technol* 2007; 16: 371–384.
38. Welsh BC, Dill NE and Zane SN. The first delinquency prevention experiment: a socio-historical review of the origins of the Cambridge-Somerville Youth Study's research design. *J Exp Criminol* 2019; 15: 441–451.
39. Cabot RC. Treatment in social case work and the need of criteria and of tests of its success and failure. *Hosp Soc Ser* 1931; 24: 435–453.
40. Thorndike EL and Ruger GJ. The effects of outside air and recirculated air upon the intellectual achievement and improvement of school pupils. *School Soc* 1916; 4: 260–264.