Journal of Experimental Criminology (2023) 19:1115–1130
https://doi.org/10.1007/s11292-022-09520-2

Pair-matching with random allocation in prospective controlled trials: the evolution of a novel design in criminology and medicine, 1926–2021

Brandon C. Welsh[1] · **Scott H. Podolsky**[2] · **Steven N. Zane**[3]

Accepted: 3 June 2022 / Published online: 18 June 2022

Abstract

Objectives Pair-matching with random allocation in prospective controlled trials represents a novel and highly rigorous design. First use of the design can be traced to medicine (in 1926) and criminology and the social sciences more generally (in 1935). Beginning with these trials, we examine the subsequent history of matched-pair RCTs (randomized controlled trials), and related attention to stratification prior to randomization, in both criminology and medicine over almost a century to illustrate shared interest in the design's advantages and disadvantages.

Methods We draw upon a wide range of historical and contemporary sources, including historical archives and writings on the first trials in criminology and medicine, prior reviews of RCTs and matched-pair RCTs, and searches of selected databases.

Results The first trials draw attention to key factors that remain central to contemporary use, including concerns about covariate imbalance when randomization is used on its own, potential to improve study power when matching is effective, and the ability to deal with differential attrition in follow-ups. The evolution of the design also shows that the single most important application of matched-pair RCTs is when the units are clusters or places.

Conclusions Over the twentieth and twenty-first centuries, criminology and medicine have continued to wrestle with methodologies to most efficiently and robustly compare like with like. Both, in this setting, have turned to matched-pair randomization, though less often than its advocates would like. It is this and other shared interests between criminology/social sciences and medicine/public health, including

✉ Brandon C. Welsh
b.welsh@northeastern.edu

[1] School of Criminology and Criminal Justice, Northeastern University, Churchill Hall, 360 Huntington Avenue, Boston, MA 02115, USA

[2] Harvard Medical School, Boston, MA, USA

[3] Florida State University, Tallahassee, FL, USA

a movement toward evidence-based policy and practice, that help us reimagine possibilities for advancing knowledge and improving public policy.

Keywords Pair-matching · Randomized controlled trial · Criminology · Medicine

Introduction

The history of randomized controlled trials (RCTs) in medicine is well known (Bothwell et al., 2016; Bothwell and Podolosky, 2016; Chalmers et al., 2012). The same can be said about the history of RCTs in assessing the effects of social interventions (Forsetlund et al., 2007), as well as more narrowly in the field of criminology (Farrington, 1983; Farrington & Welsh, 2005). Less well known is the history of another rigorous evaluation design that predated the widespread use of RCTs in criminology and medicine: pair-matching in combination with random allocation—otherwise known as matched-pair RCTs. Here, units (people or places) are matched by pairs on a wide array of covariates and units of each pair are randomly allocated to the treatment and control conditions.

Recent research identifies that this design was used as early as 1926 in medicine, and was first used in 1935 in the social sciences, specifically, criminology (Podolsky et al., 2021; Welsh et al., 2021). Within medicine, in 1926, Amberson et al. (1931) initiated a trial to investigate the efficacy of sanocrysin as a therapeutic for pulmonary tuberculosis. Twenty-four patients “free from serious complications” participated in the study. The study reports:

> 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 by two of us… 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., 1931, pp. 403-404)

In the history of RCTs in medicine, this study is considered an “outlier,” owing to alternate-allocation designs being the dominant model in the first half of the twentieth century (Bothwell & Podolsky, 2016, p. 502). Importantly, Gabriel (2014) has demonstrated the origins of the trial at the intersection of mutual public health service and pharmaceutical industry interest in an objective assessment of the drug, with the trial entailing blinding of patients to prevent a “psychic influence” on healing.

The 1935 criminology study, known as the Cambridge-Somerville Youth Study (CSYS), was initiated to evaluate the impact on youth delinquency of a social intervention of “directed friendship” (deQ Cabot, 1940). Founded and directed by Richard Clarke Cabot, a physician and professor of clinical medicine and social ethics at Harvard University, the CSYS set out to discover whether an individually focused and “morally inspired” intervention in the lives of young, disadvantaged boys could prevent them from becoming delinquent (O’Brien, 1985, p. 550). Recruitment and screening of 1953 boys, ages 5–13 years and from the cities of Cambridge and Somerville (Mass.), produced a final sample of 650. All the boys were then matched into pairs—according to 142 variables (rated on an 11-point scale)—and one member of each pair was randomly allocated, based on a coin toss, to the treatment group. There have been four assessments of delinquent and criminal behavior and other outcomes covering major periods of the life-course: transition from adolescence to early adulthood, early adulthood, middle-age, and old age (up to age 90), with the latter representing a 72-year follow-up (Welsh et al., 2019).

Cabot himself was both physician and social interventionist, and exemplified the contemporary attempt to rigorously evaluate interventions in both domains. Using Amberson et al.'s tuberculosis trial and Cabot's CSYS as the starting points, this paper examines the subsequent history of matched-pair RCTs, and related attention to stratification prior to randomization, in both medicine and criminology over almost a full century to illustrate shared interest in the advantages and disadvantages of a research design intended to ensure the comparison of like with like. Also important is consideration of implications for experimental criminology.

Background

In the 1926 sanocrysin trial, Amberson and colleagues appear to have used pair-matching followed by random allocation as a way to mitigate concerns about recruited patients presenting differing levels of symptoms of pulmonary tuberculosis. Far from just relying on the National Tuberculosis Association classification of the extent of the disease, as reported by the authors, it was "necessary to give weight to the character as well as to the extent of the disease, and also to include other clinical factors in the final judgment of the cases" (Amberson et al., 1931, p. 404). In the absence of documentation about the formal plan for the evaluation design (see Gabriel, 2014), we might infer that an equally pressing concern facing the researchers was the small number of recruited patients ($N=24$). In short, simple random allocation could not be relied upon to produce balance in the pretest measures between the treatment and control groups.

In the CSYS, Cabot in turn used pair-matching followed by random allocation because he regarded matching on its own to be insufficient. As reported by Powers and Witmer (1951, p. 78), following the matching process:

> The next question was to determine whether any given boy should fall into the treatment or the control group. It was evident that an arbitrary decision might give rise to a constant error. The proper method of determining this question was, of course, by chance. Accordingly, a coin was flipped and the cases fell into the treatment or comparison groups in accordance with its fall.

Powers and Witmer (1951, p. 78) added the following about Cabot's decision-making: "It was believed that, even if the measures used in the matching were not perfectly reliable, chance would tend to preserve, in groups as large as 325 each, an even balance of important factors."

These landmark studies draw attention to certain key factors that remain central to contemporary use of matched-pair RCTs in the social sciences and medicine. First, although random allocation is designed to help eliminate confounding, covariate imbalance is still possible. That is, the treatment and control groups may still differ by chance. This can be especially problematic in small *N* studies. Matching across known covariates can thus add "face validity" to an experimental study (Chondros et al., 2021, p. 5766).

Second, pair-matching prior to randomization can improve study power when the matching is effective, meaning that there is a positive within-pair correlation on relevant variables (Wacholder & Weinberg, 1982). By decreasing variation within matched pairs on known covariates, matching can improve the precision of estimated treatment effects compared to other designs (i.e., statistical efficiency). The relative efficiency of the matched-pair design has been demonstrated in several recent simulation studies (Balzer et al., 2015; Chondros et al., 2021), although as noted below, this will depend on the success of the matching itself (see, e.g., Ariel and Farrington (2010) on "unsuccessful blocking").

Third, randomization within matched pairs provides a straightforward way of dealing with differential attrition, which can present a serious threat to the internal validity of follow-up assessments of prospective trials. Since the proper comparison in a randomized trial involves the original treatment and control groups (i.e., "intent-to-treat"), differential attrition threatens the internal validity of the simple randomized design. Matched-pair randomization overcomes this problem since the researcher can drop both members of the pair in the event one member is missing (Farrington & Welsh, 2006). Of course, this essentially doubles the loss of follow-up, which may pose a problem for smaller studies (Ivers et al., 2012).

Perhaps the single most important application of matched-pair RCTs—and by far the dominant issue in scholarly and policy debates in both the social sciences and medicine (see Ariel & Farrington, 2010; Weisburd & Gill, 2014; Balzer et al., 2015; Imai et al., 2009a, 2009b; Chondros et al., 2021)—is when the units are clusters of individuals or places rather than individuals alone. Unlike with individual-based studies, where securing an initial *N* of some minimum threshold (e.g., 50 units in each condition; Farrington, 1983) is often straightforward, cluster- and place-based studies present any number of challenges to obtaining an initial *N* of such magnitude. Recruiting 100 or 150 schools, communities, or high-crime properties is far more difficult than obtaining a similar number of families, patients, or offenders. Small sample size is thus a key motivating factor for pair-matching in cluster- and place-based RCTs, where it appears to be most common (Campbell et al., 2007).

In the last two decades, a robust debate in medicine and public health has taken place over the potential benefits of using pair-matching in cluster-RCTs.[1] Some have gone so far as to suggest that "randomization by cluster without prior construction of matched pairs, when pairing is feasible, is an exercise in self-destruction" (Imai

[1] In medicine and public health, this design is more commonly referred to as matched-pair, cluster-randomization (or MPCR; Imai et al., 2009a). Other names for the matching component in RCTs include adaptive pair-matching and nonbipartite matching (Balzer et al., 2015). In the social sciences, a matched-pair RCT is sometimes referred to as a complete or fully blocked design. In contrast, a partially blocked design involves some type of stratifying of the cluster-based units prior to random allocation to treatment and control conditions (Weisburd & Gill 2014).

et al., 2009a, p. 48). Others have been somewhat restrained: “a randomized trial with adaptive pair-matching will often be more efficient for estimation of the CATE [conditional average treatment effect] than its completely randomized counterpart” (Balzer et al., 2015, p. 1009). Still, others have been more reserved in their enthusiasm for the design, arguing that “the actual benefits of matching in practice will not be realized unless several conditions are satisfied, conditions that may be difficult to achieve in practice” (Donner & Klar, 2004, p. 418). For example, the “degrees of freedom used to calculate the confidence interval and P-value for the intervention effect is based on the number of pairs of clusters rather than the total number of clusters,” such that pair-matching results in a substantial loss of degrees of freedom compared to simple or stratified designs (Chondros et al., 2021, p. 5766). This may pose serious problems for trials with small numbers of clusters (Donner & Klar, 2004; Ivers et al., 2012).

Most recently, Chondros and colleagues (2021) performed a simulation study comparing the efficiency of the matched-pair design with stratified and simple random designs for cluster randomized trials. The authors found that the matched-pair design was more efficient when the correlation between cluster-level outcomes within pairs was moderate to strong ($r \geq 0.3$), but not more efficient with weaker correlations.

Such deliberations have taken place alongside the evolving—if intermittent—application of a priori trial stratification and more extensive matched-pair randomization in medicine, public health, and the social sciences, as we will next illustrate.

Medicine and public health

In post-1926 prospective clinical trials in medicine, a priori matching would remain an important methodological consideration. There were those who employed matching alone, whether for ethical (Gehan & Freireich, 1974; King et al., 2006) or logistical (Inouye et al., 1999) concerns about randomization, with increasingly sophisticated measures taken to ensure the equivalence of such matching (Lin et al., 2018). However, matching alone among prospective trials appears to have been rarely practiced in the RCTs era. Rather, most discussions have focused on the relative utility of matched randomization (or before then, alternate allocation) versus randomization (or alternate allocation) alone, with discussion dating to Austin Bradford Hill’s own elaboration of the “Principles of Medical Statistics” in 1937, the same year that Cabot was enrolling his first participants in the CSYS.[2]

In Hill’s framing, it was critical in clinical trials “to ensure beforehand that, as far as is possible, the control and treated groups are the same in all *relevant* respects” (Hill, 1937a, p. 42). In alternate allocation studies, continued Hill, “*in the long run* we can fairly rely upon this random allotment of the patients to equalise in the two groups the distribution of other characteristics that may be important,” and that especially “with *large* numbers we can be reasonably sure that the numbers of each

[2] On the transition from alternate allocation studies (in which, e.g., every other patient was administered the novel remedy) to studies that entailed allocation by concealed randomization, see Chalmers (2005), Chalmers et al. (2012), and Bothwell and Podolsky (2016).

type [of differing representation with respect to particular characteristics] will be equally, or nearly equally, represented in both groups" (p. 42). However, recognizing the potential for unequal sorting in smaller studies, Hill provided a key caveat:

> If it be known that certain characteristics will have an influence upon the results of treatment and on account of relatively small numbers the distribution of these characteristics may not be equalised in the final groups, it is advisable to extend this method of allocation. For instance, alternate persons will not be treated but a division will be made by sex, so that the first male is treated and the second male untreated, the first female is treated and the second female untreated. (p. 42)

Hill later alluded to the "practical difficulties" that could enter into the design of clinical trials (Hill, 1937b). And most matched-pair randomized studies entailed only a handful of variables, with Wladyslaw Billewicz noting in 1964 that of 20 "recently published medical investigations," the number ranged from one to six, with most studies employing two or three. Debate over ensuing decades would thus focus on the relative merits and demerits of including matching prior to randomization. On the pro side of including a priori matching, a "state of 'other things being equal' is built into the design," protecting "the investigator against 'freaked' samples" (Billewicz, 1964), and, as eventually noted, improving statistical power (see, e.g., McClatchey et al., 1992).

Perhaps most prominently, in 1966, the Director of the American Medical Association's Department of Biostatistics, Stanley Schor, emphasized for *JAMA*'s audience the benefits of stratification prior to randomization: "To many clinical investigators the word 'randomization' has a magic connotation. As long as they randomize, they think it does not matter how important some pertinent characteristic is in terms of its effect on the results of a study. This may be true with enormous samples. But in the ordinary course of clinical research an investigator should not trust the randomization procedure to produce unbiased results" (Schor, 1966, p. 124). Instead, attention should be devoted early to equalize those seemingly knowable factors that could shape the trial outcomes: "If a characteristic is known to have an important effect on the experiment, an investigator should not depend upon chance in the selection process to cancel it out. The effects of important factors should be designed out of the study, controlled in some way, or allowed to remain in such a manner as to have their net effects measurable. Randomization should be relied upon only for the numerous factors of lesser importance" (Schor, 1966, p. 124). Or, as Schor concluded, the investigator "should not simply randomize and hope" (p. 124). However, statisticians were likewise willing to draw attention to the con side of the ledger, whether concerning the potentially increased cost and logistical difficulties entailed in such matching, or the potential statistical messiness it introduced (Billewicz, 1964; Bland & Altman, 1994; McKinlay, 1977).

The usage of matching within the *New England Journal of Medicine* in the twentieth and twenty-first centuries may be an instructive and representative sampling device concerning the consequent application of matching and matched-pair randomization.[3] The vast majority of "matched" investigations in the journal were retrospective case–control studies, with several hundred represented. Nonetheless, a small fraction (between 1 and 2% of the "hits" represented) were matched prospective studies. Some of these were matched, prospective observational studies: in a 1960 study of physical activity and obesity, "obese" subjects were matched by age, occupation, and socioeconomic background to "nonobese" subjects (Chirico & Stunkard, 1960), while in a 1978 study of growth and development in children with sickle-cell trait, the children were matched as closely as possible to controls according to sex, birth date, birth weight, gestational age, five-minute Apgar score, and socioeconomic status (Kramer et al., 1978). By 2015, still more elaborate methods could be used to match patients within a prospective "registry" study of patients receiving cardiac bypass surgery versus percutaneous intervention with second-generation drug-eluting stents among patients with multi-vessel coronary artery disease (Bangalore et al., 2015).

Other researchers conducted matched, prospective RCTs. The first of these, a 1961 study of vitamin C and antihistamines on gingival hyperplasia among patients receiving the anti-seizure medication phenytoin, was analogous to the study by Amberson et al. (1931), a matched-cluster randomization study (Rose et al., 1961). Later studies on the impact of glycemic control on kidney function among diabetic patients (Feldt-Rasmussen et al., 1986), and the first study of what would eventually be called copaxone for multiple sclerosis (Bornstein et al., 1987), were matched-pair studies, using three matched characteristics (albeit different ones) apiece.

Two studies, entailing matched-pair cluster randomization, shaded closer to social science investigations, with one concerning an educational program for risk factor modification for heart disease (Walter et al., 1988) and the other a safe childbirth checklist study in India (Semrau et al., 2017). That such *NEJM*-reported educational interventions noted above shared much in common with social science investigations is perhaps no surprise, given the role of biostatisticians as the shared colleagues of investigators of multiple disciplines, and the increasing ease of access of investigators across disciplines to the papers of one another (see, e.g., McKinlay, 1977). Having shown the persisting, albeit limited, application of matched-pair randomization in medicine and public health, we thus next turn to the discipline of criminology—harkening back to Cabot—and the social sciences more generally.

Criminology and the social sciences

The combination of pair-matching and random allocation in prospective controlled trials in criminology and in the social sciences is most common when the unit of allocation is clusters of individuals or places. Designs that employ some form of stratification, including pair-matching, are especially useful in this context due to

[3] Using the search term "were matched" called up 550 papers in the *NEJM* database. The first related to "matching" in a methodological sense dates from 1954, though in a review (Viets 1954). The examples described here are ones in which studies were actually conducted, rather than reported in a review article.

the smaller number of units to be allocated to treatment and control conditions. Imai et al. (2009b) reviewed pre-randomization designs in studies with cluster randomization in political science, economics, education, and medicine and public health during the 2000s. Of the 107 cluster randomized experiments that were located, 22% used stratification and 19% used pair-matching. The authors also noted that pair-matching was largely confined to studies in medicine and public health, but was also common in development economics.

Others have similarly observed that, outside of medicine and public health, cluster-RCTs with pair-matching are employed most frequently in development economics (Banerjee & Duflo, 2009). Much of this work has been conducted at MIT's Poverty Action Lab (e.g., Banerjee et al., 2007). One survey of randomized experiments in development economics found that, while most studies employed stratification prior to cluster randomization, few employed pair-matching (Bruhn & McKenzie, 2009).[4] Intriguingly, in an accompanying survey of leading researchers, approximately half indicated that they had used randomization within matched pairs at some point in their work.[5] Elsewhere, in a meta-analysis of 77 educational interventions involving random assignment procedures performed in developing countries, McEwan (2015) found that approximately 70% used some form of stratification (including pair-wise matching) prior to randomization.[6]

In the first comprehensive review of RCTs in criminology, which included published studies with a minimum $N = 100$ units (individuals or places) and covering the period 1939 to 1981, only 2 out of 37 trials used pair-matching (Farrington, 1983). One of these trials was the CSYS (McCord, 1978). The other, run by the California Youth Authority in the late 1950s, evaluated effects on recidivism of two different institutional living units (20- and 50-bed) for juvenile offenders (Jesness, 1971). Participants ($N = 281$) were matched by age and social backgrounds and then randomly allocated to either of the two treatment conditions.

An update of this review, using the same criteria and covering the period 1982–2004, identified an additional 85 RCTs (mostly of individuals) with criminological outcomes (Farrington & Welsh, 2005; see also Farrington & Welsh, 2006). Only one of the trials included pair-matching. This trial evaluated effects on recidivism of a cognitive-behavioral treatment program for male sex offenders in California (Marques et al., 1994). Participants ($N = 229$) were matched on three variables (age, prior criminal history, and offender type), arranged by pairs, and randomly allocated to either the treatment or control conditions.

[4] In this context, Bruhn and McKenzie (2009) note that stratification occurs when "units are randomly assigned to treatment and control within strata defined by usually one or two observed baseline characteristics" (p. 201), while pair-matching "provides a method to improve covariate balance for many variables at the same time" (p. 209).

[5] As described by the authors, "A notable feature of the survey responses was a much greater number of researchers randomizing within matched pairs than is apparent from the existing development literature" (Bruhn and McKenzie 2009, p. 206).

[6] The proportion of trials that used pair-matching compared to other forms of stratification was not specified.

Similar to the use of cluster randomization in the social sciences more generally, most examples of pair-matching with random allocation in criminology involve place-based experiments. Here, the unit of interest is not an individual but rather a discrete geographical area, such as a police district, high crime area ("hot spot"), business, or neighborhood (Boruch et al., 2010). In the aforementioned reviews, the included experiments with few exceptions used individuals as the unit of allocation. Since place-based experiments typically involve a small number of areas (more often < 100), pair-matching prior to random allocation provides important benefits over random allocation alone. Ideally, matched *pairs* of places could be established with one member of each pair randomly allocated to the treatment condition. This is also called a fully blocked design (Weisburd & Gill, 2014), but it is not often employed because it can entail a substantial loss of degrees of freedom (i.e., the number of variables that are free to vary following one or more restrictions placed on the data).

Policing experiments often utilize blocking prior to randomization, and on occasion, this involves pair-matching. To get a sense of the extent of the use of pair-matching in policing experiments, we drew upon the latest analysis of the Global Policing Database (GPD), as well as carried out some preliminary searches of the GPD. Developed by researchers at the University of Queensland and Queensland University of Technology in Australia, the GPD is a "web-based and searchable database designed to capture all published and unpublished experimental and quasi-experimental evaluations of policing interventions conducted since 1950" (Higginson et al., 2014; see also Eggins et al., 2016). Impressively, the GPD is updated on a fairly regular basis and it is not restricted to studies reported in English. In their latest analysis of the GPD (through 2018), Mazerolle et al. (2022) identified a total of 431 RCT of policing interventions. Based on searches of the RCTs in the database, we identified at least 20 unique studies (or 4.6%) that employed pair-matching or full blocking prior to random allocation. Some of the other RCTs used partial blocking, which involves some type of stratification of the place-based units prior to random allocation to treatment and control conditions (Weisburd & Gill, 2014).

One notable example of the use of the matched-pair RCTs design in policing was carried out by Weisburd et al. (2008) to evaluate a risk-focused policing intervention in Redlands, CA. The authors grouped 26 census blocks into 13 pairs, matched according to risk factor scores, calls for police service, population density, and median home value, and then randomly allocated units in each matched pair to receive risk-focused policing or usual patrol.

Outside of policing, there are few examples of pair-matching with random allocation in criminology. Most often, these occur in school settings where the matched-pair design is especially useful: "Since it is difficult to assign a large number of schools randomly, it may be best to place schools in matched pairs and randomly assign one member of each pair to the experimental condition and one member to the control condition" (Farrington & Ttofi, 2009, p. 327). The most notable example is Communities That Care (CTC), a multi-modal, community-based youth development program. Across seven states, 24 small, rural communities (average population = 14,646) were recruited and matched by pairs based on "population size, racial and ethnic diversity, economic indicators, and crime rates" (Hawkins et al., 2008,

p. 183). One community in each pair was then randomly assigned by coin toss to receive the preventive intervention (from grades 5 to 9). Analyses indicated baseline similarity of the intervention and control communities. Follow-up assessments have been conducted at 8 years (through grade 12; Hawkins et al., 2014) and 11 years (through age 21; Oesterle et al., 2018).

Another example of pair-matching with random allocation involved a behavioral intervention to prevent sexual assault in Nairobi, Kenya (Baiocchi et al., 2017). Thirty-two schools were pair-matched based on "number of girls in the school, number of boys in the school, academic performance, public versus private school, location, materials used to construct the school, and materials used for the floor" (Baiocchi et al., 2017, p. 822). One school from each pair was then randomly allocated to receive the intervention. Two intervention schools ultimately did not participate in the program, and the researchers dropped these schools and their matched controls. It can be concluded that the use of pair-matching in RCTs in criminology and in other social science disciplines has shown a renewed interest in the last two decades, but, like with medicine and public health, is rather limited.

Discussion and conclusions

This paper started with Cabot and Amberson to show the shared and enduring interest in criminology and medicine in rigorously comparing like with like in evaluating effects of prevention interventions and treatments. Indeed, Cabot, who was both a physician and social interventionist, showed overlap of these concerns. Over the twentieth and twenty-first centuries, both domains have continued to wrestle with methodologies to most efficiently and robustly compare like with like. Both, in this setting, have turned to pair-matching in combination with random allocation, though less often than its advocates would like.

Certainly, the boundaries can be fuzzy between criminology/social sciences and medicine/public health. Some intersection between the domains has been clearer. One important example comes from the medical profession's response to victims of violent crime. In their seminal (but non-experimental) study "Murder and Medicine," Harris et al. (2002) found that advances in emergency medical technology and care (e.g., development of 911 call systems and trauma units at hospitals, improved training for medical technicians) in the USA during the 1960s through the 1990s played a central role in increasing the chance of survival for victims of violent criminal assault. The authors estimated that the lethality of violent assaults (i.e., assaults resulting in homicides) decreased over this period of time by 2.5 to 4.5% per year.

Another notable example is the movement toward evidence-based policy and practice in the respective domains. The Cochrane Collaboration (now Cochrane) in medicine was instrumental in the founding of the Campbell Collaboration in the social sciences (which includes a major focus on crime and justice) more than 20 years ago, and the two international organizations work closely together, with many systematic reviews registered jointly (Wilson et al., 2021). Moreover, like efforts to make medicine more evidence-based, the adoption of an evidence-based approach in criminology is confronted by a number of similar obstacles, including institutional resistance and to some degree an unwillingness to learn from failures (Millenson, 2021).

Charting the evolution of this novel and highly rigorous research design in criminology and medicine over almost a full century draws attention to the possibilities for advancing knowledge and improving public policy. It also draws attention to the possibilities for experimental criminology (see Farrington et al., 2020).

Implications for experimental criminology

While most criminological research is non-experimental (Dezember et al., 2021), there has been a growing recognition that random allocation is not only necessary for establishing causal effects in evaluation research (Weisburd, 2010), but that a broad scope of criminological topics can benefit from randomized controlled trials (Ridgeway, 2019). This echoes earlier calls to make social science more experimental (Sherman, 2003), including the prediction that "[c]riminology may soon resemble medicine more than economics" (Sherman, 2005, p. 132). While criminology has not yet achieved this status (see Dezember et al., 2021), this is an intriguing observation given the historical development of pair-matching with random allocation in medicine/public health as well as in criminology.

Today, the main use of pair-matching in combination with random allocation is when the units are clusters or places, the latter often for policing interventions. In the context of a rapid growth of experimental research in criminology, as documented in the Global Policing Database and other sources (see, e.g., Farrington et al., 2020; Mazerolle et al., 2022), there are seemingly many more opportunities for researchers to use this design. Take policing, for example. Of the 431 RCTs of policing interventions in the GPD (Mazerolle et al., 2022), we identified at least 20 unique studies (or 4.6%) that used pair-matching or full blocking prior to random allocation. While this number may be small in both absolute and relative terms, it is noteworthy that most of the studies that have used this design have been conducted in the last two decades.

Understanding why some researchers who are using RCTs to evaluate police interventions are incorporating the pair-matching technique draws attention to a couple of broader themes. One has to do with the need for increased methodological rigor to achieve like with like comparisons (i.e., to improve internal validity) and increase confidence in observed effects. This takes on added importance in the context of place-based interventions when the number of units of allocation (N) is small and there is heterogeneity among the units. In this context, Weisburd and Gill (2014) demonstrate that blocking of units prior to random allocation can go a long way to decreasing covariate imbalance—and thus improving equivalence—between treatment and control conditions, without necessarily compromising statistical power or degrees of freedom. In doing so, the authors also rebut the conventional wisdom that there should be a minimum of 50 units in each condition (Farrington, 1983; Farrington & Welsh, 2006), which is not always feasible when the units are places or clusters.

Another key theme has to do with new developments in experimental methodologies and their application to criminological interventions. Most recently, Sherman (2022) reviewed the advantages and disadvantages of the repeat crossover RCT design compared to the simple (or parallel track) RCT design as applied to place-based policing interventions. In the context of the strategy of hot spots policing, Sherman (2022, p. 2) describes the repeat crossover RCT design's fundamentals:

> In this design, each hot spot serves as its own control. Using each day in each hot spot as the unit of analysis (hot spot-days), each hot spot is randomly assigned to different treatments on different days. Crime outcomes on treatment days, on average, in each hot spot are then compared to average outcomes on no-treatment days, within each hot spot.

The main advantage of this design is to allow for "continuous impact assessment" of interventions—based on "*local* knowledge"—to produce reductions in real time in targeted crimes at the local level (Sherman, 2022, p. 2, emphasis in original). Recent examples of the use of the crossover RCT design include two short duration police foot patrol interventions in hot spots of serious violence in the British city of Essex (Basford et al., 2021) and county of Bedfordshire (Bland et al., 2021).

For the criminologist designing a prospective RCT, whether it involves a simple (or parallel track), wait-list control, or some other type of design (but not crossover design), the key questions become as follows: (1) Can the units (e.g., people or places), based on the data available on the units and the recruitment process of the units, be matched into pairs prior to random allocation? and (2) Will this produce a more rigorous assessment of the impact of the intervention? The point here is that, like the principle that evaluation designs (experimental or quasi-experimental) need to be guided by the research question at-hand and not the other way around, pair-matching in combination with random allocation may not always be feasible or needed. For example, an argument could be made today that the Cambridge-Somerville Youth Study, at least based on its large original sample ($N=650$), did not require pair-matching in addition to random allocation. But, of course, this overlooks the historical context of the beginnings of experimentation in the social sciences and medicine (Bothwell & Podolsky, 2016; Forsetlund et al., 2007), not to mention concerns that Cabot had about the use of matching on its own. (Recall that for the CSYS, random allocation was a secondary consideration.) To return to the criminologist designing a prospective RCT today, even a large sample size may not be sufficient, especially if there is a moderate to high degree of heterogeneity among the units.

Whether it be through this application or others, the shared history of this particular technique for rigorously comparing like with like reinforces experimental criminology's bonds with experimentation in medicine and public health.

Acknowledgements We thank the journal editor and the anonymous reviewers for helpful comments.

References

Amberson, J. B., McMahon, B. T., & Pinner, M. (1931). A clinical trial of sanocrysin in pulmonary tuberculosis. *American Review of Tuberculosis, 24*, 401–435.

Ariel, B., & Farrington, D. P. (2010). Randomized block designs. In A. R. Piquero & D. Weisburd (Eds.), *Handbook of quantitative criminology* (pp. 437–454). Springer.

Baiocchi, M., Omondi, B., Langat, N., Boothroyd, D. B., Sinclair, J., et al. (2017). A behavior-based intervention that prevents sexual assault: The results of a matched-pairs, cluster randomized study in Nairobi, Kenya. *Prevention Science, 18*, 818–827.

Balzer, L. B., Petersen, M. L., van der Laan, M. J., the SEARCH Consortium. (2015). Adaptive pair-matching in randomized trials with unbiased and efficient effect estimation. *Statistics in Medicine, 34*, 999–1011.

Banerjee, A. V., & Duflo, E. (2009). The experimental approach to development economics. *Annual Review of Economics, 1*, 151–178.

Banerjee, A. V., Cole, S., Duflo, E., & Linden, L. (2007). Remedying education: Evidence from two randomized experiments in India. *Quarterly Journal of Economics, 122*, 1235–1264.

Bangalore, S., Guo, Y., Samadashvili, Z., Blecker, S., Xu, J., et al. (2015). Everolimus-eluting stents or bypass surgery for multivessel coronary disease. *New England Journal of Medicine, 372*, 1213–1222.

Basford, L., Sims, C., Agar, I., Harinam, V., & Strang, H. (2021). Effects of one-a-day foot patrols on hot spots of serious violence and crime harm: A randomised crossover trial. *Cambridge Journal of Evidence-Based Policing, 5*, 119–133.

Billewicz, W. Z. (1964). Matched samples in medical investigations. *British Journal of Preventive Social Medicine, 18*, 167–173.

Bland, J. M., & Altman, D. G. (1994). Matching. *British Medical Journal, 309*, 1128.

Bland, M., Leggetter, M., Cestaro, D., & Sebire, J. (2021). Fifteen minutes per day keeps the violence away: A crossover randomised controlled trial on the impact of foot patrols on serious violence in large hot spot areas. *Cambridge Journal of Evidence-Based Policing, 5*, 93–118.

Bornstein, M. B., Miller, A., Slagle, S., Weitzman, M., Crystal, H., et al. (1987). A pilot trial of cop 1 in exacerbating-relapsing multiple sclerosis. *New England Journal of Medicine, 317*, 408–414.

Boruch, R., Weisburd, D., & Berk, R. (2010). Place randomized trials. In A. R. Piquero & D. Weisburd (Eds.), *Handbook of quantitative criminology* (pp. 481–502). Springer.

Bothwell, L. E., & Podolsky, S. H. (2016). The emergence of the randomized, controlled trial. *New England Journal of Medicine, 375*, 501–504.

Bothwell, L. E., Greene, J. A., Podolsky, S. H., & Jones, D. S. (2016). Assessing the gold standard—Lessons from the history of RCTs. *New England Journal of Medicine, 374*, 2175–2181.

Bruhn, M., & McKenzie, D. (2009). In pursuit of balance: Randomization in practice in development field experiments. *American Economic Journal: Applied Economics, 1*, 200–232.

Campbell, M. J., Donner, A., & Klar, N. (2007). Developments in cluster randomized trials and *Statistics in Medicine*. *Statistics in Medicine, 26*, 2–19.

Chalmers, I. (2005). 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 G. Jorland, A. Opinel, & G. Weisz (Eds.), *Body counts: Medical quantification in historical and sociological perspectives* (pp. 309–334). McGill-Queens University Press.

Chalmers, I., Dukan, E., Podolsky, S. H., & Smith, G. D. (2012). The advent of fair treatment allocation schedules in clinical trials during the 19th and early 20th centuries. *Journal of the Royal Society of Medicine, 105*, 221–227.

Chirico, A.-M., & Stunkard, A. J. (1960). Physical activity and human obesity. *New England Journal of Medicine, 263*, 935–940.

Chondros, P., Ukoumunne, O. C., Gunn, J. M., & Carlin, J. B. (2021). When should matching be used in the design of cluster randomized trials? *Statistics in Medicine, 40*, 5765–5778.

deQ Cabot, P. S. (1940). A long-term study of children: The Cambridge-Somerville Youth Study. *Child Development, 11*, 143–151.

Dezember, A., Stoltz, M., Marmolejo, L., Kanewske, L. C., Feingold, K. D., et al. (2021). The lack of experimental research in criminology—Evidence from *Criminology* and *Justice Quarterly*. *Journal of Experimental Criminology, 17*, 677–712.

Donner, A., & Klar, N. (2004). Pitfalls of and controversies in cluster randomization trials. *American Journal of Public Health, 94*, 416–422.

Eggins, E., Higginson, A., & Mazerolle, L. (2016). Increasing the access to evidence to inform evidence-based policing: The Global Policing Database. *Police Science: Australia and New Zealand Journal of Evidence Based Policing, 1*(2), 37–38.

Farrington, D. P. (1983). Randomized experiments on crime and justice. *Crime and Justice, 4*, 257–308.

Farrington, D. P., & Ttofi, M. M. (2009). Reducing school bullying: Evidence-based implications for policy. *Crime and Justice, 38*, 281–345.

Farrington, D. P., & Welsh, B. C. (2005). Randomized experiments in criminology: What have we learned in the last two decades? *Journal of Experimental Criminology, 1*, 9–38.

Farrington, D. P., & Welsh, B. C. (2006). A half century of randomized experiments on crime and justice. *Crime and Justice, 34*, 55–132.

Farrington, D. P., Lösel, F., Braga, A. A., Mazerolle, L., Raine, A., et al. (2020). Experimental criminology: Looking back and forward on the 20th anniversary of the Academy of Experimental Criminology. *Journal of Experimental Criminology, 16*, 649–673.

Feldt-Rasmussen, B., Mathiesen, E. R., Hegedüs, L., & Deckert, T. (1986). Kidney function during 12 months of strict metabolic control in insulin-dependent diabetic patients with incipient nephropathy. *New England Journal of Medicine, 314*, 665–670.

Forsetlund, L., Chalmers, I., & Bjørndal, A. (2007). When was random allocation first used to generate comparison groups in experiments to assess the effect of social interventions? *Economics of Innovation and New Technology, 16*, 371–384.

Gabriel, J. M. (2014). The testing of sanocrysin: Science, profit, and innovation in clinical trial design, 1926–31. *Journal of the History of Medicine and Allied Sciences, 69*, 604–632.

Gehan, E. A., & Freireich, E. J. (1974). Non-randomized controls in clinical cancer trials. *New England Journal of Medicine, 290*, 198–203.

Harris, A. R., Thomas, S. H., Fisher, G. A., & Hirsch, D. J. (2002). Murder and medicine: The lethality of criminal assault 1960–1999. *Homicide Studies, 6*, 128–166.

Hawkins, J. D., Catalano, R. F., Arthur, M. W., Egan, E., Brown, E. C., et al. (2008). Testing Communities That Care: The rationale, design and behavioral baseline equivalence of the community youth development study. *Prevention Science, 9*, 178–190.

Hawkins, J. D., Oesterle, S., Brown, E. S., Abbott, R. D., & Catalano, R. F. (2014). Youth problem behaviors 8 years after implementing the Communities That Care prevention system: A community-randomized trial. *JAMA Pediatrics, 168*, 122–129.

Higginson, A., Eggins, E., Mazerolle, L., & Stanko, E. (2014). The Global Policing Database [database and protocol]. Retrieved from: http://www.gpd.uq.edu.au/

Hill, A. B. (1937). Principles of medical statistics I—The aim of the statistical method. *Lancet, 1*, 41–43.

Hill, A. B. (1937). Principles of medical statistics XV—General summary and conclusions. *Lancet, 1*, 883–885.

Imai, K., King, G., & Nall, C. (2009). The essential role of pair matching in cluster-randomized experiments, with application to the Mexican universal health insurance evaluation. *Statistical Science, 24*, 29–53.

Imai, K., King, G., & Nall, C. (2009). Rejoinder: Matched pairs and the future of cluster-randomized experiments. *Statistical Science, 24*, 65–72.

Inouye, S. K., Bogardus, S. T., Charpentier, P. A., Leo-Summers, L., Acampora, D., et al. (1999). A multicomponent intervention to prevent delirium in hospitalized older patients. *New England Journal of Medicine, 340*, 669–676.

Ivers, N. M., Halperin, I. J., Barnsley, J., Grimshaw, M., Shah, B. R., et al. (2012). Allocation techniques for balance at baseline in cluster randomized trials: A methodological review. *Trials, 13*, 1–9 (e120).

Jesness, C. F. (1971). Comparative effectiveness of two institutional treatment programs for delinquents. *Child Care Quarterly, 1*, 119–130.

King, J. C., Stoddard, J. J., Gaglani, M. J., Moore, K. A., Magder, L., et al. (2006). Effectiveness of school-based influenza vaccination. *New England Journal of Medicine, 355*, 2523–2532.

Kramer, M. S., Rooks, Y., & Pearson, H. A. (1978). Growth and development in children with sickle-cell trait. *New England Journal of Medicine, 299*, 686–689.

Lin, J., Gamao-Siebers, M., & Tiwari, R. (2018). Propensity score matched augmented controls in randomized clinical trials: A case study. *Pharmaceutical Statistics, 17*, 629–647.

Marques, J. K., Day, D. M., Nelson, C., & West, M. A. (1994). Effects of cognitive-behavioral treatment on sex offender recidivism: Preliminary results of a longitudinal study. *Criminal Justice and Behavior, 21*, 28–54.

Mazerolle, L., Eggins, E., Hine, L., & Higginson, A. (2022). The role of randomized experiments in developing the evidence for evidence-based policing. In D. Weisburd, T. Jonathan-Zamir, B. Hasisi, & G. Perry (Eds.), *The future of evidence-based policing*. Cambridge University Press, in press.

McClatchey, M. W., Cohen, S. J., & Reed, F. M. (1992). The usefulness of matched pair randomization for medical practice-based research. *Family Practice Research Journal, 12*, 235–243.

McCord, J. (1978). A thirty-year follow-up of treatment effects. *American Psychologist, 33*, 284–289.

McEwan, P. J. (2015). Improving learning in primary schools of developing countries: A meta-analysis of randomized experiments. *Review of Educational Research, 85*, 353–394.

McKinlay, S. M. (1977). Pair-matching—A reappraisal of a popular technique. *Biometrics, 33*, 725–735.

Millenson, M. L. (2021). Docs and cops: Origins and ongoing challenges of evidence-based policing. *Cambridge Journal of Evidence-Based Policing, 5*, 146–155.

O'Brien, L. (1985). A bold plunge into the sea of values': The career of Dr. Richard Cabot. *New England Quarterly, 58*, 533–553.

Oesterle, S., Kuklinski, M. R., Hawkins, J. D., Skinner, M. L., Guttmannova, K., & Rhew, I. C. (2018). Long-term effects of the Communities That Care trial on substance use, antisocial behavior, and violence through age 21 years. *American Journal of Public Health, 108*, 659–665.

Podolsky, S. H., Welsh, B. C., & Zane, S. N. (2021). Richard Cabot, pair-matched random allocation, and the attempts to compare like with like in the social sciences and medicine. Part 2: The context of medicine and public health. *Journal of the Royal Society of Medicine, 114*, 264–270.

Powers, E., & Witmer, H. L. (1951). *An experiment in the prevention of delinquency: The Cambridge-Somerville Youth Study*. Columbia University Press.

Ridgeway, G. (2019). Experiments in criminology: Improving our understanding of crime and the criminal justice system. *Annual Review of Statistics and Its Application, 6*, 37–61.

Rose, J. H., Buffaloe, W. J., & Dunham, R. M. (1961). A study of the effect of vitamin C and of an antihistaminic drug on gingival hyperplasia occurring in patients receiving diphenylhydantoin sodium. *New England Journal of Medicine, 265*, 932–935.

Sanchez, E., Robertson, T. R., Lewis, C. M., Rosenbluth, B., Bohman, T., et al. (2001). Preventing bullying and sexual harassment in elementary schools. *Journal of Emotional Abuse, 2*, 157–180.

Schor, S. S. (1966). The mystic statistic: Randomization. *Journal of the American Medical Association, 196*, 124.

Semrau, K. E. A., Hirschhorn, L. R., Delaney, M. M., Singh, V. P., Saurastri, R., et al. (2017). Outcomes of a coaching-based WHO safe child birth checklist program in India. *New England Journal of Medicine, 377*, 2313–2324.

Sherman, L. W. (2003). Misleading evidence and evidence-led policy: Making social science more experimental. *Annals of the American Academy of Political and Social Science, 589*, 6–19.

Sherman, L. W. (2005). The use and usefulness of criminology, 1751–2005: Enlightened justice and its failures. *Annals of the American Academy of Political and Social Science, 600*, 115–135.

Sherman, L. W. (2022). "Test-as-you-go" for hot spots policing: Continuous impact assessment with repeat crossover deigns. *Cambridge Journal of Evidence-Based Policing*.https://doi.org/10.1007/s41887-022-00073-y

Viets, H. R. (1954). Myasthenia gravis. *New England Journal of Medicine, 251*, 97–105.

Wacholder, S., & Weinberg, C. R. (1982). Paired versus two-sample design for a clinical trial of treatments with dichotomous outcome: Power considerations. *Biometrics, 38*, 801–812.

Walter, H. J., Hofman, A., Vaughan, R. D., & Wynder, E. L. (1988). Modification of risk factors for coronary heart disease. *New England Journal of Medicine, 318*, 1093–1100.

Weisburd, D. (2010). Justifying the use of non-experimental methods and disqualifying the use of randomized controlled trials: Challenging the folklore in evaluation research in crime and justice. *Journal of Experimental Criminology, 6*, 209–227.

Weisburd, D., & Gill, C. (2014). Block randomized trials at places: Rethinking the limitations of small N experiments. *Journal of Quantitative Criminology, 30*, 97–112.

Weisburd, D., Morris, N. A., & Ready, J. (2008). Risk-focused policing at places: An experimental evaluation. *Justice Quarterly, 25*, 163–200.

Welsh, B. C., Podolsky, S. H., & Zane, S. N. (2021). Richard Cabot, pair-matched random allocation, and the attempt to compare like with like in the social sciences and medicine. Part 1: The context of the social sciences. *Journal of the Royal Society of Medicine, 114*, 212–217.

Welsh, B. C., Zane, S. N., Zimmerman, G. M., & Yohros, A. (2019). Association of a crime prevention program for boys with mortality 72 years after the intervention: A follow-up of a randomized clinical trial. *JAMA Network Open, 2*, 1–11 (e190782).

Wilson, D. B., Mazerolle, L., & Neyroud, P. (2021). Campbell Collaboration systematic reviews and the Journal of Experimental Criminology: Reflections on the last 20 years. *Journal of Experimental Criminology, 17*, 539–544.

Brandon C. Welsh Ph.D., is Professor of Criminology at Northeastern University, Director of the Cambridge-Somerville Youth Study, and, starting July 1, 2022, the Visiting Professor of Global Health and Social Medicine at Harvard Medical School. His research focuses on the prevention of delinquency, crime, and violence and evidence-based social policy. He has written extensively on these topics and is the author or editor of 12 books, including, with Steven Zane and Daniel Mears, *The Oxford Handbook of Evidence-Based Crime and Justice Policy* (Oxford University Press, forthcoming). He received a Ph.D. from Cambridge University.

Scott H. Podolsky M.D., is Professor of Global Health and Social Medicine at Harvard Medical School, a primary care physician at Massachusetts General Hospital, and the Director of the Center for the History of Medicine at the Francis A. Countway Library of Medicine. He has written extensively on the history of controlled clinical trials, especially in two of his monographs, *Pneumonia before Antibiotics: Therapeutic Evolution and Evaluation in Twentieth-Century America* (Johns Hopkins University Press, 2006) and *The Antibiotic Era: Reform, Resistance, and the Pursuit of a Rational Therapeutics* (Johns Hopkins University Press, 2015). He received an M.D. from Harvard Medical School.

Steven N. Zane Ph.D., J.D., is an Assistant Professor of Criminology at Florida State University. His research interests focus on crime over the life-course, juvenile justice, and evidence-based social policy. He is the author or editor of two books, including *Explaining Variation in Juvenile Punishment: The Role of Communities and Systems* (Routledge Press, 2021), and an author of numerous scientific journal articles and book chapters. He received a Ph.D. from Northeastern University and a J.D. from Boston College Law School.