World Journal of Methodology

Submit a Manuscript: https://www.f6publishing.com

World J Methodol 2024 March 20; 14(1): 90590

DOI: 10.5662/wjm.v14.i1.90590

ISSN 2222-0682 (online)

MINIREVIEWS

Can propensity score matching replace randomized controlled trials?

Matthias Yi Quan Liau, En Qi Toh, Shamir Muhamed, Surya Varma Selvakumar, Vishalkumar Girishchandra Shelat

Specialty type: Methodology

Provenance and peer review: Invited article; Externally peer reviewed.

Peer-review model: Single blind

Peer-review report's scientific quality classification

Grade A (Excellent): 0
Grade B (Very good): 0
Grade C (Good): C, C
Grade D (Fair): 0
Grade E (Poor): 0

P-Reviewer: Nooripour R, Iran

Received: December 7, 2023
Peer-review started: December 7, 2023
First decision: December 23, 2023
Revised: January 5, 2024
Accepted: February 23, 2024
Article in press: February 23, 2024
Published online: March 20, 2024

Matthias Yi Quan Liau, En Qi Toh, Shamir Muhamed, Surya Varma Selvakumar, Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore 308232, Singapore

Vishalkumar Girishchandra Shelat, Department of General Surgery, Tan Tock Seng Hospital, Singapore 308433, Singapore

Vishalkumar Girishchandra Shelat, Surgical Science Training Centre, Tan Tock Seng Hospital, Singapore 308433, Singapore

Corresponding author: Vishalkumar Girishchandra Shelat, FEBS, FRCS, MBBS, MMed, Adjunct Associate Professor, Department of General Surgery, Tan Tock Seng Hospital, 11 Jalan Tan Tock Seng, Singapore 308433, Singapore. vgshelat@gmail.com

Abstract

Randomized controlled trials (RCTs) have long been recognized as the gold standard for establishing causal relationships in clinical research. Despite that, various limitations of RCTs prevent its widespread implementation, ranging from the ethicality of withholding potentially-lifesaving treatment from a group to relatively poor external validity due to stringent inclusion criteria, amongst others. However, with the introduction of propensity score matching (PSM) as a retrospective statistical tool, new frontiers in establishing causation in clinical research were opened up. PSM predicts treatment effects using observational data from existing sources such as registries or electronic health records, to create a matched sample of participants who received or did not receive the intervention based on their propensity scores, which takes into account characteristics such as age, gender and comorbidities. Given its retrospective nature and its use of observational data from existing sources, PSM circumvents the aforementioned ethical issues faced by RCTs. Majority of RCTs exclude elderly, pregnant women and young children; thus, evidence of therapy efficacy is rarely proven by robust clinical research for this population. On the other hand, by matching study patient characteristics to that of the population of interest, including the elderly, pregnant women and young children, PSM allows for generalization of results to the wider population and hence greatly increases the external validity. Instead of replacing RCTs with PSM, the synergistic integration of PSM into RCTs stands to provide better research outcomes with both methods complementing each other. For example, in an RCT investigating the impact of mannitol on outcomes among participants of the Intensive Blood Pressure Reduction in Acute Cerebral

Hemorrhage Trial, the baseline characteristics of comorbidities and current medications between treatment and control arms were significantly different despite the randomization protocol. Therefore, PSM was incorporated in its analysis to create samples from the treatment and control arms that were matched in terms of these baseline characteristics, thus providing a fairer comparison for the impact of mannitol. This literature review reports the applications, advantages, and considerations of using PSM with RCTs, illustrating its utility in refining randomization, improving external validity, and accounting for non-compliance to protocol. Future research should consider integrating the use of PSM in RCTs to better generalize outcomes to target populations for clinical practice and thereby benefit a wider range of patients, while maintaining the robustness of randomization offered by RCTs.

Key Words: Propensity score matching; Randomized controlled trials; Randomization; Clinical practice; Validity; Ethics

Core Tip: Several studies in the literature compare treatment effect estimates in propensity score matching studies and randomized controlled trials (RCTs), but few employ both methods synergistically in determining treatment outcomes. This is a first review to report and provide examples on how propensity score matching can be integrated into RCTs to refine randomization, account for non-compliance to protocol and improve external validity to produce more comprehensive and generalizable evidence for informed clinical decision making.

Citation: Liau MYQ, Toh EQ, Muhamed S, Selvakumar SV, Shelat VG. Can propensity score matching replace randomized controlled trials? *World J Methodol* 2024; 14(1): 90590
URL: https://www.wjgnet.com/2222-0682/full/v14/i1/90590.htm
DOI: https://dx.doi.org/10.5662/wjm.v14.i1.90590

INTRODUCTION

In the paradigm of clinical research, establishing causality is vital in helping clinicians better grasp the efficacy, or harm, of potentially groundbreaking interventions. The foundation for evidence-based medicine depends on the influence of interventions on patients' health. To attain the best outcomes for patients, causal relationships must be studied objectively. In this regard, randomized controlled trials (RCTs) have been well-established as the gold standard for establishing causal relationships in clinical research, largely due to the randomization of participants which eliminates confounders[1]. Despite this, there still exist concerns about the use of RCTs in clinical research. This includes the ethicality of withholding potentially-lifesaving treatment from a group, relatively poor external validity of RCTs due to stringent inclusion criteria, need for resources to conduct trials, and lack of feasibility to continue trials for a prolonged duration due to manpower and resource constraints[2]. In particular, a retrospective cohort study of 1017 RCTs found that poor recruitment is the most frequently reported reason for RCT termination, accounting for up to 40% of the total discontinuations[3]. This in turn leads to a considerable waste of scarce research resources. However, with the inception of propensity score matching (PSM) as a retrospective statistical methodology by Rosenbaum and Rubin[4] in 1983, new frontiers have opened to establish causality in clinical research. PSM predicts treatment effects using observational data from registries to create a matched sample of participants who received or did not receive the intervention based on their propensity scores, which is associated with baseline characteristics such as age. Since then, newer models of computing propensity scores have emerged and the uptake of PSM in research has increased exponentially, owing to its ability to estimate causal effects when random assignment of treatments is unethical or not feasible[5-9]. Recently, an increasing number of studies have begun to adopt an integrated approach, increasing the generalizability of their results with PSM while maintaining the robustness of randomization offered by RCTs, demonstrating the potential benefits of using both methods synergistically[10-12]. Therefore, this literature review aims to explore the integration of PSM studies as a potent adjunct to RCTs in establishing causality in healthcare, potentially circumventing the concerns and quandaries surrounding RCTs as a research modality. We report the advantages and limitations of RCTs and PSM studies, as well as their synergistic implementation and its advantages as compared to either method alone. It is hoped that the findings of our review would help guide researchers and clinicians alike to consider adopting the use of PSM in RCTs in future clinical research.

METHODOLOGY

A comprehensive literature search was conducted on PubMed, Web of Science, CENTRAL, Scopus and Embase from inception to 25 November 2023 with the keywords: "randomized controlled trial", "propensity score matching", "observational studies", and "advantages". The inclusion criteria included studies with PSM or RCTs as their methodology, or if both methods were used concurrently. Non-English articles, animal studies, conference abstracts, oral

and poster presentations were excluded. If an institution published more than one study, the most recent article was selected for analysis. Relevant information regarding the advantages and limitations of RCTs, PSM and their synergistic implementation were extracted. The included articles also served as key examples to support the use of each method or the integration of PSM into RCTs respectively.

RANDOMIZED CONTROLLED TRIALS

An RCT is an experimental study randomly assigning human participants to a control and treatment group, and is typically used to establish cause and effect in medical treatments such as novel drugs, devices and surgical techniques. Crucially, the randomization process ensures that participants are allocated to groups comparable in terms of baseline characteristics and potential confounders, and that the observed differences in outcomes are due to the treatment effect [13,14]. different types and methods of randomization are performed in RCTs. Some common methods are shown in Table 1[15,16]. The choice of randomization method depends on the characteristics and objectives of the study, such as the sample size, the number of treatment groups, the presence of covariates, and the primary outcome. The randomization method should be specified in the study protocol and implemented with adequate concealment to ensure the validity and reliability of the results[17].

In double blinded studies, blinding is also incorporated to ensure that neither patients nor doctors administering the treatment are aware of the treatment allocation[18,19]. In triple blinded studies, data analysts assessing the outcomes are also blinded to reduce bias further. This is crucial as knowledge of which treatment the patient receives could lead to the behavioral changes of patients and doctors who might be biased towards the provision of the newer treatment instead of the placebo, decreasing the objectivity and credibility of the study[20]. Blinding can be achieved through the certain means including double-dummy designs, central randomization and independent outcome assessors. Furthermore, it is recommended to use pre-specified and transparent protocols, registration, and reporting guidelines to ensure the integrity of the blinding process. Despite that, some studies cannot be blinded[21]. For example, blinding cannot be achieved in a trial comparing different types of psychotherapy for depression as the patients and therapists would know which type of therapy they are receiving or providing[22]. Overall, randomization and blinding seek to reduce allocation and selection biases within RCTs. Figure 1A summarizes the major steps taken in an RCT study.

PROPENSITY SCORE MATCHING

An alternative to estimating causal relationships when RCTs cannot be performed is PSM. PSM is a statistical technique that predicts treatment or interventional effects using observational data from existing sources such as registries or electronic health records, to create a matched sample of participants who received or did not receive the intervention based on their propensity scores[4]. Propensity scores are the probabilities of receiving the intervention given the observed characteristics of the participants such as age, gender and comorbidities, and attempts to reduce the bias and confounders inherently present in such studies[23]. This is due to participants being assigned treatments and interventions based on clinical needs, mostly influenced by patient-centered factors including but not limited to patient comorbidities, and not through random allocation[4]. For example, in a study comparing the effectiveness of a new drug *vs* a placebo, the participants who choose to take the new drug may differ from those who do not in terms of their age, health status, or other characteristics that may affect the outcome of interest. These differences may confound the causal relationship between the treatment and the outcome, and make the comparison between the groups invalid.

Various models are used to compute such probabilities, which are outlined in Table 2. The logistic regression model is most commonly used due to its simplicity. However, it assumes a linear relationship between the covariates and the log-odds of the treatment, an assumption that does not always hold true for all variables in medicine[24]. Body mass index (BMI), for instance, has a nonlinear relationship with mortality, with both low and high BMI being associated with a high mortality rate[25]. In such cases, other models can be used assuming the variables fulfill the model's assumptions[26]. It is important to note that different models used will generate different results in a finite sample. Although many models exist, there are no guidelines for their choice. However, one could consider the guidelines proposed by Baser[27] based on five criteria which aims to select the best model based on their ability to reduce selection bias in that given data sample.

PSM then involves matching the participants who received the intervention with those who did not based on their propensity scores, so that they are theorized to have similar characteristics and would thus be comparable. PSM can use different matching methods depending on the availability and balance of the data[28]. Table 3 shows the possible matching methods available while Figure 1B summarizes the major steps taken in a study employing PSM.

CAN PROPENSITY SCORE MATCHING REPLACE RANDOMIZED CONTROLLED TRIALS?

Although PSM and RCTs are used to establish a causal link between interventions and eventual health outcomes, it is important to note the differences between these methodologies. Fundamentally, RCTs are conducted to determine the efficacy of new treatments compared to an existing control treatment[29]. In contrast, PSM studies serve as a retrospective method of evaluating the efficacy of a treatment. This is done by using the baseline characteristics of patients to match individualized treatment groups against newly formed control groups *via* assigning a propensity score to each group[30].

Table 1 Comparison of randomization methods for clinical trials

Method	Description	Ref.
Simple randomization	Each participant has an equal chance of being assigned to any of the treatment groups. This method is easy to implement and unpredictable, but it may result in unequal group sizes or imbalances in important covariates, especially in small studies	Grimm and Müller[75], 1999
Block randomization	Participants are allocated to treatment groups in blocks of fixed size, such as 4 or 6. This method ensures that the group sizes are balanced at any point of the study, but it may introduce some predictability if the block size is known or guessed by the investigators	Sreedevi *et al* [76], 2017
Stratified randomization	Participants are first stratified by one or more relevant factors, such as age, gender, or disease severity, and then randomized within each stratum. This method ensures that the treatment groups are balanced with respect to the stratification factors, but it may increase the complexity and cost of the randomization process	Kahan and Morris[21], 2012
Minimization	Participants are allocated to the treatment group that minimizes the imbalance in a set of predefined factors, such as prognostic variables or previous treatments. This method is adaptive and can achieve better balance than stratified randomization, but it may also introduce some predictability and bias if the allocation is not concealed	Treasure and MacRae[77], 1998

Table 2 Comparison of methods of computing propensity scores

Method	Advantages	Disadvantages	Ref.
Logistic regression	Simple and widely used	May not capture complex or nonlinear relationships	Otok *et al*[6], 2017
	Can handle binary and continuous covariates	May be sensitive to model misspecification	
	Can estimate the propensity score and the treatment effect in one model	May not balance all covariates well	
Discriminant analysis	Can handle multiclass treatment	May not capture nonlinear relationships	Rudner and Johnette [7], 2006
	Can capture linear combinations of covariates	May be sensitive to outliers and distributional assumptions	
	Can handle multicollinearity among covariates	May not balance all covariates well	
Random forests	Can handle complex and nonlinear relationships	May be computationally intensive	Zhao *et al*[8], 2016
	Can handle binary, categorical, and continuous covariates	May overfit the data	
	Can balance all covariates well	May not estimate the propensity score and the treatment effect in one model	

Table 3 Possible matching methods utilized in propensity score matching studies

Matching method	Indication
One-to-one	This method matches each treated unit with one control unit that has the closest propensity score. This method is simple and intuitive, but it may discard some units that are not matched
One-to-many	This method matches each treated unit with more than one control unit that has similar propensity scores. This method can increase the sample size and precision, but it may also introduce more bias due to imperfect matches
Nearest neighbor	This method matches each treated unit with the control unit that has the nearest propensity score, within a specified caliper or threshold. This method can reduce bias by excluding poor matches, but it may also reduce efficiency by excluding good matches
Caliper	This method matches each treated unit with the control unit that has the propensity score within a specified range or distance. This method can ensure a high degree of similarity between the matched pairs, but it may also result in a loss of observations if the caliper is too narrow
Stratification	This method divides the propensity score distribution into a number of strata or intervals, and then compares the outcomes of the treated and control units within each stratum. This method can balance the covariates across the strata, but it may also produce heterogeneous treatment effects across the strata

As will be explored further in the paper, specific advantages and disadvantages to both study methodologies would favour clinicians and researchers to choose one method over the other (Table 4). However, it is also important to contextualize the considerations behind the methodology choice[31].

Context and objectives

By the nature of RCTs (Figure 1A), the division of participants into two distinct groups, encompassing one control group and one treatment group, necessitates certain ethical considerations, particularly concerning the control group and their

Table 4 Summary of the advantages of propensity score matching and randomized controlled trials

Propensity score matching	RCTs
Allows for utilization of retrospective data where randomization was not done	Gold standard for causal inference by eliminating bias
Improves efficiency of subject enrolment in prospective studies	Required as part of regulatory requirements
Allows analysis of causal inference in investigations where ethical considerations forbid RCTs	Allows researchers to conduct targeted studies to answer specific questions
Better external validity and generalizability	Better internal validity
Avoidance of type II errors	
Shorter timeline to study completion	

RCTs: Randomized controlled trials.

access to a potential novel treatment or intervention. Firstly, due to the presence of strict inclusion and exclusion criteria that governs those who are eligible for participation within an RCT, there is a chance that certain populations or groups may be consistently excluded from taking part in these studies that would determine the efficacy of potentially life-saving treatments. Consequently, the results obtained from the study cannot be universalizable, as the effects of the treatment and intervention cannot be accurately assessed in these ignored groups[32].

This can be observed in studies such as the one done by Leinonen *et al*[33], which demonstrates a discrepancy in the biodata of RCT subjects as compared to the general population, particularly concerning the use of acetylcholinesterase inhibitors as a treatment for Alzheimer's disease. The study found that RCT participants were significantly younger, due to the stringency of exclusion criteria that prevented the recruitment of older individuals into the study, who usually comprise the bulk of the demographic who are usually afflicted with Alzheimer's disease. As such, the use of acetylcholinesterase inhibitors as a treatment option in older adults cannot be fully validated. Were PSM to be used instead, it would follow that the generalizability of the results would allow for findings to be applied across various age groups, including older adults, who would likely benefit the most from treatment. Further, strict selection criteria of RCT might lead to more favorable treatment outcomes for one group compared to another group. For example, Gui *et al*[34] reported that surgical resection had superior 3-year and 5-year disease free survival compared to transarterial chemoembolization plus radiofrequency ablation for hepatocellular carcinoma. When analyzing only the PSM data, the difference was not significant.

Moreover, there is the need to consider the ethicality of using placebos, which is often done in the control group in RCTs. Especially in the presence of established first-line treatment or intervention options for whichever disease or condition is being studied, the use of placebos in place of this established treatment option would be ethically frowned upon, more so if the condition being studied is potentially deadly or debilitating and swift, definitive intervention would be necessary[35,36]. Due to the retrospective nature of PSM studies, which use already-existing data to determine the efficacy of treatments, it is possible to eliminate these ethical concerns as the study design circumvents the need to recruit participants to assess interventional outcomes[37]. A key example is examining the cardiovascular safety of common oral hypoglycemic agents for type 2 diabetes mellitus. As the study aimed to investigate the adverse effects of drug therapies, it would be unethical to expose patients to the risk of harm in RCTs. Instead, an algorithm is developed based on generalized PSM to estimate the effects of various diabetic medications individually or with metformin on cardiovascular events[38]. A case could be made for PSM studies replacing RCTs in such instances.

Despite the advantages that PSM has over the ethical considerations of RCTs, it is also important not to discount the ethical considerations of PSM itself, particularly regarding data privacy and the use of patient data from electronic health records and registries. As it is often not practical to seek the consent of individuals to utilize observational data for retrospective studies, institutional review boards usually waive the need to obtain consent for using personal data in most circumstances. This is provided that the results are published in the form that does not identify the individual. Despite that, the issue of data privacy in the face of the retrospective nature of PSM studies must not be ignored. Nevertheless, the advent of methods such as propensity score-based pooling and combining distributed linear regression with propensity score modelling can avoid the need for individual-level data while maintaining analytic integrity, thereby offering protection of patient privacy[39,40]. Hence, PSM could still prove to be the better method when ethical issues arise.

Feasibility

Although RCTs are known for their robustness in determining the efficacy of new treatments, they are also known to take a longer time to complete, especially due to the 'real-time' nature of tracking participant progress and long-term outcomes. The entire RCT progress, as well as its duration, is dependent on the recruitment of appropriate participants for the study. This process requires the development of strict criteria that would enable only the most suitable subjects to participate in the study. The development of such criteria, together with the time it takes to recruit sufficient suitable participants to participate in the study, would understandably take a long time[41]. In addition to this, one must consider that participants must be followed up for an extended period because treatments themselves have several potential outcomes and side effects that may take months or even years to manifest. Additional time would also need to be taken to

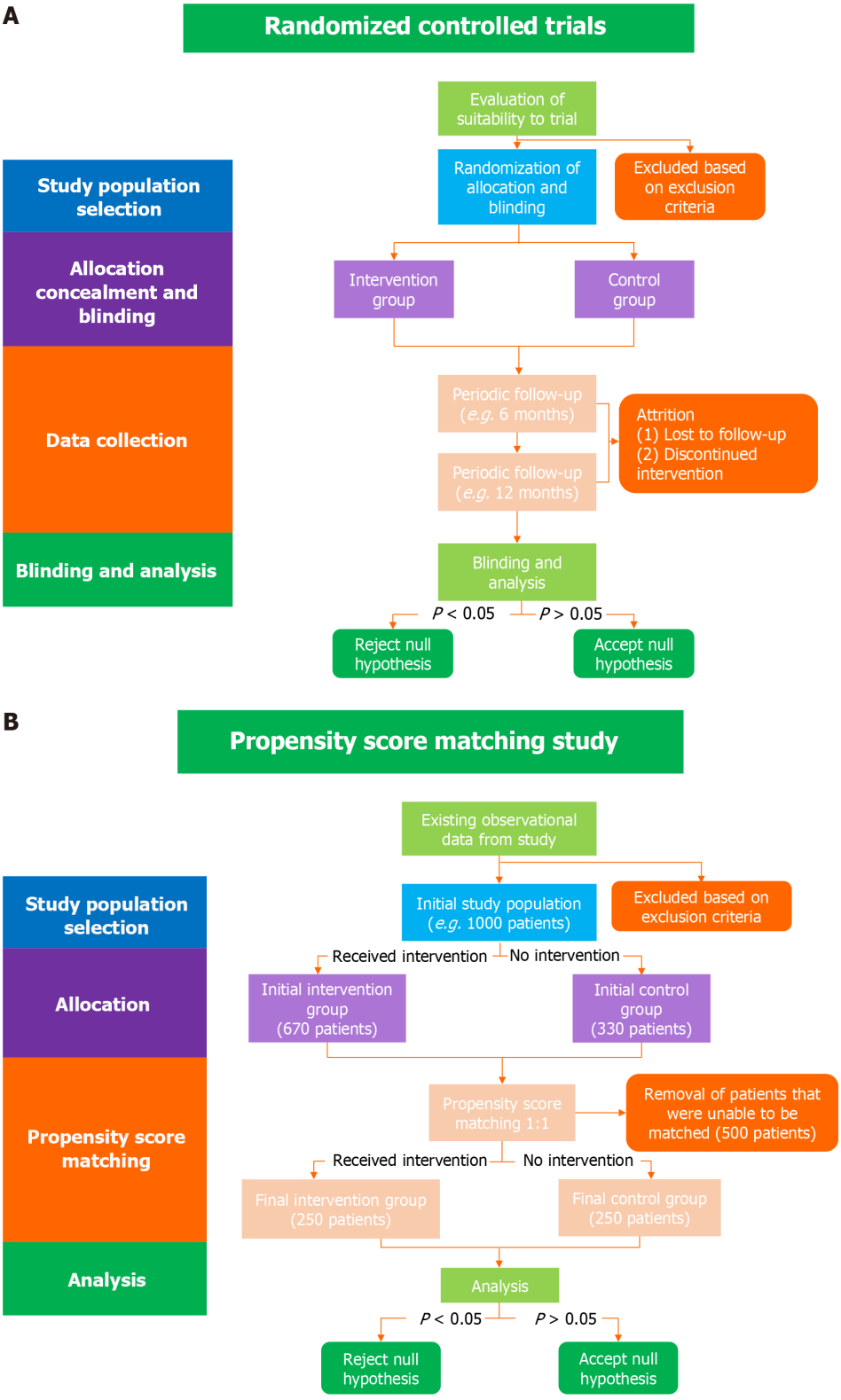

Figure 1 Flowchart overviewing the steps. A: Flowchart overviewing the steps taken in a randomized controlled trial, randomization of allocation and blinding is key to reduce bias in randomized controlled trials; B: Flowchart overviewing the steps taken in a propensity score matching study, propensity score matching creates a matched sample of patients who received or did not receive the intervention based on their propensity scores, which takes into account the patients' baseline characteristics.

obtain data of interest and perform statistical analysis to form conclusions regarding treatment efficacy[42]. Majority of research grants have an expiry date of couple of years by which a RCT has to be concluded. Thus, it is a common observation that recurrence free survival (RFS) is commonly used as a surrogate of overall survival (OS) in oncology trials as RFS can be determined much earlier as compared to OS.

On the other hand, PSM studies can be conducted on pre-existing data and would take much shorter time to complete. In instances where a treatment for a condition would be required with some urgency, such as in the case of vaccines or antidotes for epidemics and pandemics, it is possible that PSM studies could replace RCTs. Consider this study by Hsu *et al*[43], which uses a PSM methodology with data from a previous cohort study to suggest new influenza vaccination guidelines for the elderly. Especially given the morbidity of the infection in older populations, as well as the dynamism of the infection itself due to high mutation rates and the existence of multiple strains of the same virus, vaccination recommendations would have to be generated rather quickly to adapt to an ever-changing seasonal infection. This was possible due to PSM, whose methodology allowed for retrospective comparison between vaccinated and non-vaccinated individuals, and the effects of vaccination specifically on those vaccinated. Therefore, PSM studies are usually more feasible and quicker to implement due to their retrospective nature which precludes the need for recruitment and monitoring.

Better external validity

As stringent prerequisites are used in RCTs regarding participant selection, a strict set of inclusion and exclusion criteria are often needed. This, verily, would have implications for the validity of RCTs. Such exacting criteria would mean that RCTs generally have poorer external validity, which means that study findings may not be generalizable to the rest of the population, or different contexts[44]. Majority of RCTs exclude elderly, pregnant women and young children; thus, evidence of therapy efficacy is rarely proven by robust clinical research for this population. On the other hand, the validity of PSM studies depends on how closely the study sample represents the population of concern. This is ultimately influenced by multiple factors, including the study context, confounders, and the statistical model used to yield the actual propensity score. It is generally accepted that the external validity of PSM studies is quite robust, as the results from PSM studies can be generalized to other populations given that the model assumptions are accounted for. The sample size is sufficient enough for a large statistical power[45]. Studies exist that show that, at the very least, PSM provides an external validity that is comparable, if not better, than that of RCTs. For example, a study by Kuss *et al*[46] compared PSM and RCT as methodologies for assessing outcomes following coronary artery bypass grafting. The study validated that any differences observed in the findings between the RCT and PSM methodologies in this specific context were statistically insignificant. Thus, owing to better external validity, PSM can be employed to better generalize outcomes of studies to patient populations for translation to clinical practice.

Addressing lack of randomization in retrospective studies

Retrospective studies are an important area where the advantages of PSM are demonstrated. They use existing data recorded for purposes other than research and patients usually do not undergo interventions *via* randomization[47]. These data are thus best analyzed *via* PSM. For example, in a retrospective cohort comparison study of cervical total disc replacement performed as an outpatient *vs* inpatient procedure, patients were not randomized to either outpatient or inpatient groups unlike in RCTs. Pre-existing variables thus likely influenced the type of intervention they received. To eliminate their influence, PSM was used to control for variances in patient characteristics. Every patient was assigned a propensity score based on variables such as age and gender, among others. Each outpatient case was then systematically matched with three inpatient cases with similar propensity scores to compare intervention outcomes[48]. After adjusting for the inherent confounding factors, any differences in observed outcomes can be attributed to the intervention itself although randomization is not performed[28]. Therefore, PSM can make use of retrospective data in a way to analyze the causal effects of the intervention itself.

The advent of PSM also allows better application of findings generated from retrospective analyses. Although robust, RCTs are challenging to conduct and often generate results that may not apply to a real-world setting. This may be due to either the complexity of the intervention or the selection process for participants yielding a population different from that seen in general clinical practice[49]. Unlike RCTs, PSM makes use of data that have been collected from actual patients undergoing interventions in real-world practices. This gives the analysis a more realistic touch and makes it more applicable to clinical practice.

Resource efficiency

PSM has its role in prospective cohort studies as well. PSM can be used for patient enrolment in prospective studies to improve statistical and logistical efficiency. In a novel approach to PSM, a propensity score model is developed based on pre-existing patient data. The study tapped on data from two groups of patients-those who were referred for acupuncture and those receiving the usual care, to compare the effectiveness of the two interventions to manage chronic musculoskeletal pain. These patients are not randomized to either group. Patients are then matched by their propensity scores for recruitment into the prospective cohort study, where patient-reported outcomes are collected through an interview. Without PSM, patients that would have otherwise been recruited would ultimately be excluded from analysis due to a lack of propensity score overlap[50]. This thus improves study precision and maximizes resources.

Another example where PSM can be used to increase efficiency in terms of patient enrolment is the Diabetes Prevention Program (DPP)[51]. DPP is a multicenter RCT designed to compare diet and exercise against medications on preventing or delaying the onset of type 2 diabetes. However, the process of subject selection was highly inefficient. The 158177 subjects had to be screened before 3819 subjects were finally randomized to one of the four original arms. If PSM had been employed to recruit the subjects in a more targeted fashion, less resources could have been expended.

Error avoidance

With the use of PSM, it may be possible to avoid the type II error that often affects the statistical power of RCTs. RCTs

have a high risk of type II error, failing to reject the null hypothesis when it is false[52]. In other words, RCTs may falsely report no significant difference between the intervention and the control groups. This may be due to the rigorous nature of RCTs, which require careful planning, ethical approval, recruitment, randomization, intervention delivery, follow-up, data collection, analysis, and reporting. These processes can introduce limitations that can reduce the power and precision of RCTs, such as low sample size, high attrition, poor adherence, crossover, contamination, protocol deviations, and measurement error. Despite using various protocols such as intention-to-treat, as-treated, or per-protocol analysis, RCTs may still fail to detect or report clinically significant changes in the outcome of interest[44]. Conversely, PSM utilizes existing data sets, thus circumventing the issues related to low statistical power that plague RCTs, especially in small sample sizes and high attrition rates.

WHEN ARE RANDOMIZED CONTROLLED TRIALS BETTER?

RCTs have been the acknowledged standard in evidence-based medicine for decades, only second to systematic reviews and meta-analyses[53]. The performance of a RCT is robust and requires strict specification of study conditions in all aspects of its conduct, including participant selection, treatment and control assignment arms, inclusion and exclusion criteria, randomization method, outcome measurement, among others[49]. It first emerged in 1948 to investigate streptomycin treatment of pulmonary tuberculosis[54] and is recognized as the standard method for "rational therapeutics" in medicine by the 20th century[55]. Although novel methods of proving causal effects have emerged, RCTs are still highly regarded due to their various advantages which will be discussed below and summarized in Table 4.

Elimination of bias: Gold standard for causal inference

RCTs have the unique advantage of randomization which eliminates accidental bias, including selection bias. This adjusts for inherent features that may have increased the likelihood of subjects being allocated to treatment or control groups. Randomization thus eliminates any systematic differences between the two groups. This promotes comparability of the study groups, creating a basis for more accurate comparison[56], which has not been possible in other study designs[1]. As a result, any outcome differences can be attributed to the intervention rather than confounding factors. This contributes to the high internal validity of RCTs as a study design. To top it all off, RCTs can provide high statistical power, detecting and quantifying meaningful effect size differences between the intervention and the control groups, proving the causal relationship between intervention and outcomes more robustly[57].

While observational studies may use statistical methods to try to account for possible bias, some biases are very hard to correct[58]. A 2020 systematic review by Lantz[59] of 46 evaluations of interventions targeted at healthcare super-utilizers warned of this caveat. Methodological and study design weaknesses, especially regression to the mean, called into question supposed positive findings. Interestingly, observational studies of super-utilizer programs tended to report positive outcomes post-intervention. Yet on the other hand, RCTs reported no significant difference between intervention and control groups. The "positive" outcomes of these observational studies were likely biased by regression to the mean. This refers to the statistical tendency for patients incurring unusually high costs at a particular point in time to move closer to the average over time[60]. Despite statistical methods in place to correct for bias, this may not always be successful, depending on the inherent features of the data set. Therefore, this further strengthens the gold standard status of RCTs for causal inference.

While PSM can correct for confounders, it still has its shortcomings especially when compared to RCTs. Most importantly, it assumes that all relevant confounders are measured and included in the propensity score model. This is known as the ignorability or unconfoundedness assumption, and it is often untestable and may be violated in practice [61]. If unobserved or unmeasured confounders are present but not accounted for by the propensity scores, the matching may not eliminate them. Consequently, causal estimates may be biased or inconsistent[62]. Therefore, without a careful selection and measurement of the covariates based on substantive knowledge and theory, PSM may remain inferior to RCTs.

Although PSM is a thorough process, the possibility of bias due to matching errors or model misspecification cannot be overlooked. PSM estimates propensity scores with a statistical model, such as logistic regression, discriminant analysis, or random forests. These models may be misspecified or inaccurate, thus not capturing the true relationship between the covariates and the treatment assignment[63]. The process may produce mismatched pairs with poorly estimated propensity scores, increasing the comparison's imbalance or bias[64]. Therefore, RCTs may retain their role as the gold standard for causal inference, until such systematic shortcomings in the alternatives are accounted for.

Regulatory requirements

Given its status as the gold standard for causal inference, RCTs have a long-standing role in regulatory requirements. Since 1962, in the wake of the thalidomide crisis in which an anti-nausea and sedative drug widely used was found to cause severe congenital disabilities, evidence of efficacy is required before a drug can be approved[65]. The Food and Drug Administration (FDA) in the United States is a key player in the approval of drugs and medical devices[66]. Under FDA regulations, a series of clinical trials are conducted with the medication, to determine if the findings support the manufacturer's efficacy claims and demonstrate that the drug is safe. The early drug approval statute in 1962 generally required at least two adequate and well-controlled randomized investigations[67]. Although regulatory guidelines have evolved over the past decades to allow non-RCTs as well as to include a range of concessions, RCTs continue to have a long-lasting importance in this field, given their rigor and advantage of randomization.

Targeted studies

A clear difference between RCTs and observational studies is that when ethical and feasible, RCTs allow researchers to design a study to investigate questions they want answered, rather than the questions they can answer with naturally occurring data. For instance, the impact of physician-patient race concordance on patient's health behavior is notoriously challenging to determine using observational data[68]. As most individuals choose their primary care physician, selection is already present in concordant *vs* discordant dyads. Further, long-standing structural inequalities have it such that many disadvantaged individuals, who tend to be of a minority race, do not have a primary care physician[58]. If researchers were to rely on existing data through observational studies alone, this question would never be answered sufficiently. In contrast, RCTs have the capacity to provide a satisfactory response to this. Researchers created a pop-up clinic where Black male patients were randomly assigned to see either a Black or non-Black physician[69]. It was found that those randomly assigned to Black physicians were 18% more likely to use preventive health services after the interaction than those assigned to a racially discordant doctor. This example thus illustrates an added advantage of RCTs in addressing questions that cannot otherwise be answered with regular observational studies.

USE OF PROPENSITY SCORE MATCHING IN RANDOMIZED CONTROLLED TRIALS

While RCTs are the gold standard for causal inference, challenges such as ethical considerations and real-world applicability can limit the scope and generalizability of RCT findings. This is when PSM can complement RCTs, offering a solution to address these issues and enhance the validity of the results. The synergy between RCTs and PSM can be powerful, and several examples illustrate how these two methodologies can work together effectively.

RCTs with imperfect randomization

In some RCTs, the process of randomization may not achieve perfect balance in baseline covariates, especially in small samples. This can lead to residual confounding and affect the internal validity of the trial. PSM can be employed further to improve the balance between the treatment and control groups, enhancing the reliability of the RCT results. For example, the Bracing in Adolescent Idiopathic Scoliosis Trial was initially designed solely as a randomized trial[70]. However, there was a slower than anticipated enrollment of participants due to most participants preferring one treatment over the other and thus declining randomization. Therefore, a preference cohort was included thereafter and PSM was used to control for potential selection bias due to the nonrandom treatment assignment in the preference cohort. This helped to refine the treatment and control arms in both the randomized and preference cohorts in terms of baseline characteristics, and facilitates integration of the data for fair comparison. In another study by Wang *et al*[71] investigating the impact of mannitol on outcome among participants of the Intensive Blood Pressure Reduction in Acute Cerebral Hemorrhage Trial (INTERACT2), it was found that there was significant variability in baseline covariates between patients treated with and without mannitol. With the use of propensity score methods, the baseline characteristics of both cohorts can be adjusted for, and a fairer comparison can be made to determine the effect of mannitol. Van Groenestijn *et al* [72] also used PSM to correct for baseline inequalities in a RCT studying the effectiveness of aerobic exercise therapy on disease-specific and generic health-related quality of life in ambulatory patients with amyotrophic lateral sclerosis. Hence, the use of propensity score analyses and multivariate models can be synergistic with RCTs to establish causal relationships while enhancing validity.

Non-compliance to protocol

In RCTs, there are often participants who do not adhere to the assigned treatment. In such scenarios, PSM can be applied to account for non-compliance or deviations in treatment received, and thus provide a more comprehensive understanding of the treatment's impact. The Odyssey Outcomes trial compared the cardiovascular outcomes of treatment using alirocumab with placebo in patients with recent acute coronary syndrome receiving intensive statin treatment[73]. Despite being prescribed specific doses of alirocumab, some patients did not adhere to the prescribed dose or frequency of medication, which would affect the perceived effectiveness of the drug. To account for this, PSM was used to adjust for patients' compliance to the prescribed drug regime so that a better comparison between alirocumab and placebo could be performed. Therefore, propensity score methods may be useful in accounting for nonadherence or deviations in protocol which may be inevitable in large clinical studies.

Translation of RCT data to clinical practice

RCTs are designed to establish causal relationships under controlled conditions, but the extrapolation of their results to broader patient populations and diverse clinical settings poses several challenges, such as the use of stringent inclusion and exclusion criteria to enhance internal validity while compromising external validity, limiting the generalizability of findings to real-world patient populations. In addition, clinical practice involves diverse patient populations with varying comorbidities, demographics, and treatment responses. RCTs may not capture this heterogeneity adequately[74]. Paradoxically, many clinical practices are rapidly adopted by medical practitioners despite no evidence from RCTs. For example, laparoscopic cholecystectomy, a current gold-standard procedure for removing gallbladders, is not supported by RCT evidence.

PSM addresses the challenges of translating RCT data to clinical practice by facilitating a more nuanced comparison of treatment and control groups. By accounting for observed confounding variables, PSM helps create matched cohorts that closely resemble the characteristics of the broader patient population encountered in clinical settings. For example, RCTs

have demonstrated that the Songling Xuemaikang capsule (SXC) is effective in reducing blood pressure in essential hypertension. However, the efficacy of SXC in actual clinical settings is still unknown. Using a PSM approach, Lai *et al*[10] compared the results of patients treated with SXC monotherapy from both real world and RCT cohorts and found that SXC monotherapy is at least as effective in real-world settings as within the RCT. Similarly, Chung *et al*[11] used propensity score-based poststratification to generalize the results of the Flexibility in Duty Hour Requirements for Surgical Trainees Trial to the nonrepresentative samples. In addition, Godley *et al*[12] used propensity score based methods to assess the impact of dosage levels of Volunteer Recovery Support for Adolescents across measures such as frequency of substance used and emotional problems. Therefore, PSM employed using data from RCTs to create a matched cohort reflective of the broader patient population can allow for a more realistic assessment of the intervention's effectiveness in routine clinical settings.

LIMITATIONS

There are some limitations to our review. The studies reviewed in this article were not able to do a direct comparison between PSM and RCTs due to the nature of their investigation. Nevertheless, these studies were able to describe the advantages and disadvantages of each method collectively. In addition, there is a lack in reporting of the disadvantages of integrating PSM into RCTs. Further studies are therefore required to examine the limitations of the synergistic implementation of PSM and RCTs concurrently.

CONCLUSION

More studies adopting the synergistic implementation of PSM and RCTs concurrently are emerging, demonstrating the feasibility and advantages the integration of both methods have to offer. PSM offers an ethical and practical alternative in situations where RCTs are not feasible or ethical. RCTs, on the other hand, continue to be the gold standard for establishing causal relationships, offering the highest level of internal validity and have a role in regulatory requirement for novel medical treatment. Ultimately, the choice between PSM and RCTs should be made carefully, considering the specific goals and constraints of the research context applied. Rather than a binary choice, the integration of PSM into RCTs should also be considered if possible. The combined implementation of both approaches can help improve the generalizability of results to a wider range of patients and specific patient populations of interest for translation to clinical practice, while maintaining the robustness of randomization and high internal validity. Therefore, the synergistic integration of PSM into RCTs should be considered for future research when possible.

FOOTNOTES

Author contributions: Liau MYQ, Toh EQ, Muhamed S, Selvakumar SV and Shelat VG designed the research study; Liau MYQ, Toh EQ, Muhamed S, Selvakumar SV and Shelat VG performed the research; Liau MYQ, Toh EQ, Muhamed S, Selvakumar SV and Shelat VG analyzed the data and wrote the manuscript; All authors have read and approve the final manuscript.

Conflict-of-interest statement: All the authors report no relevant conflicts of interest for this article.

Country/Territory of origin: Singapore

ORCID number: Matthias Yi Quan Liau 0009-0007-0623-2477; En Qi Toh 0000-0002-8469-936X; Vishalkumar Girishchandra Shelat 0000-0003-3988-8142.

S-Editor: Li L
L-Editor: A
P-Editor: Zhao S

REFERENCES

1 **Hariton E**, Locascio JJ. Randomised controlled trials - the gold standard for effectiveness research: Study design: randomised controlled trials. *BJOG* 2018; **125**: 1716 [PMID: 29916205 DOI: 10.1111/1471-0528.15199]
2 **Matthews JNS**. Introduction to randomized controlled clinical trials. 2nd ed. New York: CRC Press, 2006 [DOI: 10.1201/9781420011302]
3 **Kasenda B**, von Elm E, You J, Blümle A, Tomonaga Y, Saccilotto R, Amstutz A, Bengough T, Meerpohl JJ, Stegert M, Tikkinen KA,

Neumann I, Carrasco-Labra A, Faulhaber M, Mulla SM, Mertz D, Akl EA, Bassler D, Busse JW, Ferreira-González I, Lamontagne F, Nordmann A, Gloy V, Raatz H, Moja L, Rosenthal R, Ebrahim S, Schandelmaier S, Xin S, Vandvik PO, Johnston BC, Walter MA, Burnand B, Schwenkglenks M, Hemkens LG, Bucher HC, Guyatt GH, Briel M. Prevalence, characteristics, and publication of discontinued randomized trials. *JAMA* 2014; **311**: 1045-1051 [PMID: 24618966 DOI: 10.1001/jama.2014.1361]

4 **Rosenbaum PR**, Rubin DB. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* 1983; **70**: 41-55 [DOI: 10.1093/biomet/70.1.41]

5 **Stürmer T**, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *J Clin Epidemiol* 2006; **59**: 437-447 [PMID: 16632131 DOI: 10.1016/j.jclinepi.2005.07.004]

6 **Otok BW**, Aisyah A, Purhadi, Andari S. Propensity score matching of the gymnastics for diabetes mellitus using logistic regression. *AIP Conference Proceedings* 2017; **1913** [DOI: 10.1063/1.5016668]

7 **Rudner LM**, Johnette P. Consider propensity scores to compare treatments. *Appl Med Artif Intell (2022)* 2006; **11**: 9 [DOI: 10.7275/kqk5-wa12]

8 **Zhao P**, Su X, Ge T, Fan J. Propensity score and proximity matching using random forest. *Contemp Clin Trials* 2016; **47**: 85-92 [PMID: 26706666 DOI: 10.1016/j.cct.2015.12.012]

9 **Rosenbaum PR**, Rubin DB. Propensity scores in the design of observational studies for causal effects. *Biometrika* 2022; **110**: 1-13 [DOI: 10.1093/biomet/asac054]

10 **Lai X**, Fang Z, Dong Z, Wu S, Zhou X, Gao Y. A propensity score matched comparison of blood pressure lowering in essential hypertension patients treated with antihypertensive Chinese herbal Medicine: comparing the real-world registry data vs. randomized controlled trial. *Clin Exp Hypertens* 2023; **45**: 2249269 [PMID: 37639695 DOI: 10.1080/10641963.2023.2249269]

11 **Chung JW**, Bilimoria KY, Stulberg JJ, Quinn CM, Hedges LV. Estimation of Population Average Treatment Effects in the FIRST Trial: Application of a Propensity Score-Based Stratification Approach. *Health Serv Res* 2018; **53**: 2567-2590 [PMID: 28833067 DOI: 10.1111/1475-6773.12752]

12 **Godley MD**, Passetti LL, Hunter BD, Griffin BA. Volunteer Recovery Support for Adolescents: Using propensity score based methods to understand dosage effects within a randomized controlled trial. *J Subst Abuse Treat* 2022; **132**: 108637 [PMID: 34654584 DOI: 10.1016/j.jsat.2021.108637]

13 **Galbraith S**, Bowden J, Mander A. Accelerated longitudinal designs: An overview of modelling, power, costs and handling missing data. *Stat Methods Med Res* 2017; **26**: 374-398 [PMID: 25147228 DOI: 10.1177/0962280214547150]

14 **Ferreira JC**, Patino CM. Randomization: beyond tossing a coin. *J Bras Pneumol* 2016; **42**: 310 [PMID: 27812628 DOI: 10.1590/S1806-37562016000000296]

15 **Dettori J**. The random allocation process: two things you need to know. *Evid Based Spine Care J* 2010; **1**: 7-9 [PMID: 22956922 DOI: 10.1055/s-0030-1267062]

16 **Kang M**, Ragan BG, Park JH. Issues in outcomes research: an overview of randomization techniques for clinical trials. *J Athl Train* 2008; **43**: 215-221 [PMID: 18345348 DOI: 10.4085/1062-6050-43.2.215]

17 **Berger VW**, Bour LJ, Carter K, Chipman JJ, Everett CC, Heussen N, Hewitt C, Hilgers RD, Luo YA, Renteria J, Ryeznik Y, Sverdlov O, Uschner D; Randomization Innovative Design Scientific Working Group. A roadmap to using randomization in clinical trials. *BMC Med Res Methodol* 2021; **21**: 168 [PMID: 34399696 DOI: 10.1186/s12874-021-01303-z]

18 **Forbes D**. Blinding: an essential component in decreasing risk of bias in experimental designs. *Evid Based Nurs* 2013; **16**: 70-71 [PMID: 23696228 DOI: 10.1136/eb-2013-101382]

19 **Karanicolas PJ**, Farrokhyar F, Bhandari M. Practical tips for surgical research: blinding: who, what, when, why, how? *Can J Surg* 2010; **53**: 345-348 [PMID: 20858381]

20 **Ot'alora G M**, Grigsby J, Poulter B, Van Derveer JW 3rd, Giron SG, Jerome L, Feduccia AA, Hamilton S, Yazar-Klosinski B, Emerson A, Mithoefer MC, Doblin R. 3,4-Methylenedioxymethamphetamine-assisted psychotherapy for treatment of chronic posttraumatic stress disorder: A randomized phase 2 controlled trial. *J Psychopharmacol* 2018; **32**: 1295-1307 [PMID: 30371148 DOI: 10.1177/0269881118806297]

21 **Kahan BC**, Morris TP. Reporting and analysis of trials using stratified randomisation in leading medical journals: review and reanalysis. *BMJ* 2012; **345**: e5840 [PMID: 22983531 DOI: 10.1136/bmj.e5840]

22 **Sil A**, Kumar P, Kumar R, Das NK. Selection of Control, Randomization, Blinding, and Allocation Concealment. *Indian Dermatol Online J* 2019; **10**: 601-605 [PMID: 31544090 DOI: 10.4103/idoj.IDOJ_149_19]

23 **Austin PC**. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behav Res* 2011; **46**: 399-424 [PMID: 21818162 DOI: 10.1080/00273171.2011.568786]

24 **Rosenblum M**, van der Laan MJ. Using regression models to analyze randomized trials: asymptotically valid hypothesis tests despite incorrectly specified models. *Biometrics* 2009; **65**: 937-945 [PMID: 19210739 DOI: 10.1111/j.1541-0420.2008.01177.x]

25 **Liu Y**, Hu H, Han Y, Li L, Li Z, Zhang L, Luo Z, Huang G, Lan Z. Body Mass Index Has a Nonlinear Association With Postoperative 30-Day Mortality in Patients Undergoing Craniotomy for Tumors in Men: An Analysis of Data From the ACS NSQIP Database. *Front Endocrinol (Lausanne)* 2022; **13**: 868968 [PMID: 35518931 DOI: 10.3389/fendo.2022.868968]

26 **Garcia Iglesias D**. Propensity Score Matching Underestimates Real Treatment Effect, in a Simulated Theoretical Multivariate Model. *Mathematics* 2022; **10**: 1547 [DOI: 10.3390/math10091547]

27 **Baser O**. Too much ado about propensity score models? Comparing methods of propensity score matching. *Value Health* 2006; **9**: 377-385 [PMID: 17076868 DOI: 10.1111/j.1524-4733.2006.00130.x]

28 **Weisburd D**, Wilson DB, Wooditch A, Britt C. Propensity Score Matching. In: Weisburd D, Wilson DB, Wooditch A, Britt C, editors. Advanced Statistics in Criminology and Criminal Justice. Cham: Springer International Publishing, 2022: 417-449 [DOI: 10.1007/978-3-030-67738-1_10]

29 **Stanley K**. Design of randomized controlled trials. *Circulation* 2007; **115**: 1164-1169 [PMID: 17339574 DOI: 10.1161/CIRCULATIONAHA.105.594945]

30 **Jupiter DC**. Propensity Score Matching: Retrospective Randomization? *J Foot Ankle Surg* 2017; **56**: 417-420 [PMID: 28231973 DOI: 10.1053/j.jfas.2017.01.013]

31 **Kuss O**, Blettner M, Börgermann J. Propensity Score: an Alternative Method of Analyzing Treatment Effects. *Dtsch Arztebl Int* 2016; **113**: 597-603 [PMID: 27658473 DOI: 10.3238/arztebl.2016.0597]

32 **Fortin M**, Dionne J, Pinho G, Gignac J, Almirall J, Lapointe L. Randomized controlled trials: do they have external validity for patients with

multiple comorbidities? *Ann Fam Med* 2006; **4**: 104-108 [PMID: 16569712 DOI: 10.1370/afm.516]

33 **Leinonen A**, Koponen M, Hartikainen S. Systematic Review: Representativeness of Participants in RCTs of Acetylcholinesterase Inhibitors. *PLoS One* 2015; **10**: e0124500 [PMID: 25933023 DOI: 10.1371/journal.pone.0124500]

34 **Gui CH**, Baey S, D'cruz RT, Shelat VG. Trans-arterial chemoembolization + radiofrequency ablation *vs* surgical resection in hepatocellular carcinoma - A meta-analysis. *Eur J Surg Oncol* 2020; **46**: 763-771 [PMID: 31937433 DOI: 10.1016/j.ejso.2020.01.004]

35 **Miller FG**, Brody H. What makes placebo-controlled trials unethical? *Am J Bioeth* 2002; **2**: 3-9 [PMID: 12189059 DOI: 10.1162/152651602317533523]

36 **Solomon P**, Cavanaugh MM, Draine J, Solomon P. Ethical Considerations of Randomized Controlled Trials. In: Solomon P, Cavanaugh MM, Draine J. Randomized Controlled Trials: Design and Implementation for Community-Based Psychosocial Interventions. Oxford: Oxford University Press, 2009 [DOI: 10.1093/acprof:oso/9780195333190.003.0002]

37 **Monlezun DJ**, Sinyavskiy O, Peters N, Steigner L, Aksamit T, Girault MI, Garcia A, Gallagher C, Iliescu C. Artificial Intelligence-Augmented Propensity Score, Cost Effectiveness and Computational Ethical Analysis of Cardiac Arrest and Active Cancer with Novel Mortality Predictive Score. *Medicina (Kaunas)* 2022; **58** [PMID: 36013506 DOI: 10.3390/medicina58081039]

38 **Scotina AD**, Zullo AR, Smith RJ, Gutman R. Approximate Bayesian Bootstrap procedures to estimate multilevel treatment effects in observational studies with application to type 2 diabetes treatment regimens. *Stat Methods Med Res* 2020; **29**: 3362-3380 [PMID: 32588747 DOI: 10.1177/0962280220928109]

39 **Rassen JA**, Solomon DH, Curtis JR, Herrinton L, Schneeweiss S. Privacy-maintaining propensity score-based pooling of multiple databases applied to a study of biologics. *Med Care* 2010; **48**: S83-S89 [PMID: 20473213 DOI: 10.1097/MLR.0b013e3181d59541]

40 **Toh S**, Wellman R, Coley RY, Horgan C, Sturtevant J, Moyneur E, Janning C, Pardee R, Coleman KJ, Arterburn D, McTigue K, Anau J, Cook AJ. Combining distributed regression and propensity scores: a doubly privacy-protecting analytic method for multicenter research. *Clin Epidemiol* 2018; **10**: 1773-1786 [PMID: 30568510 DOI: 10.2147/CLEP.S178163]

41 **Zabor EC**, Kaizer AM, Hobbs BP. Randomized Controlled Trials. *Chest* 2020; **158**: S79-S87 [PMID: 32658656 DOI: 10.1016/j.chest.2020.03.013]

42 **Llewellyn-Bennett R**, Bowman L, Bulbulia R. Post-trial follow-up methodology in large randomized controlled trials: a systematic review protocol. *Syst Rev* 2016; **5**: 214 [PMID: 27978859 DOI: 10.1186/s13643-016-0393-3]

43 **Hsu PS**, Lian IB, Chao DY. A Population-Based Propensity Score-Matched Study to Assess the Impact of Repeated Vaccination on Vaccine Effectiveness for Influenza-Associated Hospitalization Among the Elderly. *Clin Interv Aging* 2020; **15**: 301-312 [PMID: 32184579 DOI: 10.2147/CIA.S238786]

44 **Freiman JA**, Chalmers TC, Smith HA, Kuebler RR. The importance of beta, the type II error, and sample size in the design and interpretation of the randomized controlled trial: survey of two sets of "negative" trials. In: Bailar III JC, Mosteller F. Medical uses of statistics. Boca Raton: CRC Press, 2019: 357-389 [DOI: 10.1201/9780429187445-19]

45 **Stuart EA**, Cole SR, Bradshaw CP, Leaf PJ. The use of propensity scores to assess the generalizability of results from randomized trials. *J R Stat Soc Ser A Stat Soc* 2001; **174**: 369-386 [PMID: 24926156 DOI: 10.1111/j.1467-985X.2010.00673.x]

46 **Kuss O**, Legler T, Börgermann J. Treatments effects from randomized trials and propensity score analyses were similar in similar populations in an example from cardiac surgery. *J Clin Epidemiol* 2011; **64**: 1076-1084 [PMID: 21482068 DOI: 10.1016/j.jclinepi.2011.01.005]

47 **Hess DR**. Retrospective studies and chart reviews. *Respir Care* 2004; **49**: 1171-1174 [PMID: 15447798]

48 **Bovonratwet P**, Fu MC, Tyagi V, Ondeck NT, Albert TJ, Grauer JN. Safety of Outpatient Single-level Cervical Total Disc Replacement: A Propensity-Matched Multi-institutional Study. *Spine (Phila Pa 1976)* 2019; **44**: E530-E538 [PMID: 30247372 DOI: 10.1097/BRS.0000000000002884]

49 **Boyko EJ**. Observational research--opportunities and limitations. *J Diabetes Complications* 2013; **27**: 642-648 [PMID: 24055326 DOI: 10.1016/j.jdiacomp.2013.07.007]

50 **Johnson ES**, Dickerson JF, Vollmer WM, Rowley AM, Ritenbaugh C, Deyo RA, DeBar L. The feasibility of matching on a propensity score for acupuncture in a prospective cohort study of patients with chronic pain. *BMC Med Res Methodol* 2017; **17**: 42 [PMID: 28302054 DOI: 10.1186/s12874-017-0318-4]

51 **Rubin RR**, Fujimoto WY, Marrero DG, Brenneman T, Charleston JB, Edelstein SL, Fisher EB, Jordan R, Knowler WC, Lichterman LC, Prince M, Rowe PM; DPP Research Group. The Diabetes Prevention Program: recruitment methods and results. *Control Clin Trials* 2002; **23**: 157-171 [PMID: 11943442 DOI: 10.1016/S0197-2456(01)00184-2]

52 **Lochner HV**, Bhandari M, Tornetta P 3rd. Type-II error rates (beta errors) of randomized trials in orthopaedic trauma. *J Bone Joint Surg Am* 2001; **83**: 1650-1655 [PMID: 11701786 DOI: 10.2106/00004623-200111000-00005]

53 **Wolfenden L**, Foy R, Presseau J, Grimshaw JM, Ivers NM, Powell BJ, Taljaard M, Wiggers J, Sutherland R, Nathan N, Williams CM, Kingsland M, Milat A, Hodder RK, Yoong SL. Designing and undertaking randomised implementation trials: guide for researchers. *BMJ* 2021; **372**: m3721 [PMID: 33461967 DOI: 10.1136/bmj.m3721]

54 **Streptomycin in Tuber-culosis Trials Committee**. Streptomycin treatment of pulmonary tuberculosis. *Br Med J* 1948; **2**: 769-782 [DOI: 10.1136/bmj.2.4582.769]

55 **Bondemark L**, Ruf S. Randomized controlled trial: the gold standard or an unobtainable fallacy? *Eur J Orthod* 2015; **37**: 457-461 [PMID: 26136438 DOI: 10.1093/ejo/cjv046]

56 **Lim CY**, In J. Randomization in clinical studies. *Korean J Anesthesiol* 2019; **72**: 221-232 [PMID: 30929415 DOI: 10.4097/kja.19049]

57 **Walach H**. Old habits die hard--a reply to Adam La Caze. *J Clin Epidemiol* 2016; **72**: 7-9 [PMID: 26836254 DOI: 10.1016/j.jclinepi.2016.01.024]

58 **Alsan M**, Finkelstein AN. Beyond Causality: Additional Benefits of Randomized Controlled Trials for Improving Health Care Delivery. *Milbank Q* 2021; **99**: 864-881 [PMID: 34288117 DOI: 10.1111/1468-0009.12521]

59 **Lantz PM**. "Super-Utilizer" Interventions: What They Reveal About Evaluation Research, Wishful Thinking, and Health Equity. *Milbank Q* 2020; **98**: 31-34 [PMID: 32030820 DOI: 10.1111/1468-0009.12449]

60 **Barnett AG**, van der Pols JC, Dobson AJ. Regression to the mean: what it is and how to deal with it. *Int J Epidemiol* 2005; **34**: 215-220 [PMID: 15333621 DOI: 10.1093/ije/dyh299]

61 **King G**, Nielsen R. Why Propensity Scores Should Not Be Used for Matching. *Political Analysis* 2019; **27**: 435-454 [DOI: 10.1017/pan.2019.11]

62 **Shadish W**, Cook TD, Campbell DT. Experimental and quasi-experimental designs for generalized causal inference. NY: Houghton Mifflin Company, 2002

63 **Pearl J**. Causality: Models, Reasoning and Inference. 2nd ed. Cambridge: Cambridge University Press, 2009 [DOI: 10.1017/CBO9780511803161]

64 **Chan KS**, Junnarkar SP, Low JK, Huey CWT, Shelat VG. Aging is Associated with Prolonged Hospitalisation Stay in Pyogenic Liver Abscess-A 1:1 Propensity Score Matched Study in Elderly Versus Non-Elderly Patients. *Malays J Med Sci* 2022; **29**: 59-73 [PMID: 36474543 DOI: 10.21315/mjms2022.29.5.7]

65 **Kim JH**, Scialli AR. Thalidomide: the tragedy of birth defects and the effective treatment of disease. *Toxicol Sci* 2011; **122**: 1-6 [PMID: 21507989 DOI: 10.1093/toxsci/kfr088]

66 **Keng MK**, Wenzell CM, Sekeres MA. A drug's life: the pathway to drug approval. *Clin Adv Hematol Oncol* 2013; **11**: 646-655 [PMID: 24518374]

67 **Darrow JJ**, Avorn J, Kesselheim AS. FDA Approval and Regulation of Pharmaceuticals, 1983-2018. *JAMA* 2020; **323**: 164-176 [PMID: 31935033 DOI: 10.1001/jama.2019.20288]

68 **Shen MJ**, Peterson EB, Costas-Muñiz R, Hernandez MH, Jewell ST, Matsoukas K, Bylund CL. The Effects of Race and Racial Concordance on Patient-Physician Communication: A Systematic Review of the Literature. *J Racial Ethn Health Disparities* 2018; **5**: 117-140 [PMID: 28275996 DOI: 10.1007/s40615-017-0350-4]

69 **Alsan M**, Garrick O, Graziani G. Does Diversity Matter for Health? Experimental Evidence from Oakland. *American Economic Review* 2019; **109**: 4071-4111 [DOI: 10.1257/aer.20181446]

70 **Weinstein SL**, Dolan LA, Wright JG, Dobbs MB. Effects of bracing in adolescents with idiopathic scoliosis. *N Engl J Med* 2013; **369**: 1512-1521 [PMID: 24047455 DOI: 10.1056/NEJMoa1307337]

71 **Wang X**, Arima H, Yang J, Zhang S, Wu G, Woodward M, Muñoz-Venturelli P, Lavados PM, Stapf C, Robinson T, Heeley E, Delcourt C, Lindley RI, Parsons M, Chalmers J, Anderson CS; INTERACT2 Investigators. Mannitol and Outcome in Intracerebral Hemorrhage: Propensity Score and Multivariable Intensive Blood Pressure Reduction in Acute Cerebral Hemorrhage Trial 2 Results. *Stroke* 2015; **46**: 2762-2767 [PMID: 26265125 DOI: 10.1161/STROKEAHA.115.009357]

72 **van Groenestijn AC**, Schröder CD, van Eijk RPA, Veldink JH, Kruitwagen-van Reenen ET, Groothuis JT, Grupstra HF, Tepper M, van Vliet RO, Visser-Meily JMA, van den Berg LH. Aerobic Exercise Therapy in Ambulatory Patients With ALS: A Randomized Controlled Trial. *Neurorehabil Neural Repair* 2019; **33**: 153-164 [PMID: 30744529 DOI: 10.1177/1545968319826051]

73 **Schwartz GG**, Gabriel Steg P, Bhatt DL, Bittner VA, Diaz R, Goodman SG, Jukema JW, Kim YU, Li QH, Manvelian G, Pordy R, Sourdille T, White HD, Szarek M; Odyssey Outcomes Committees and Investigators. Clinical Efficacy and Safety of Alirocumab After Acute Coronary Syndrome According to Achieved Level of Low-Density Lipoprotein Cholesterol: A Propensity Score-Matched Analysis of the Odyssey Outcomes Trial. *Circulation* 2021; **143**: 1109-1122 [PMID: 33438437 DOI: 10.1161/CIRCULATIONAHA.120.049447]

74 **Rothwell PM**. External validity of randomised controlled trials: "to whom do the results of this trial apply?". *Lancet* 2005; **365**: 82-93 [PMID: 15639683 DOI: 10.1016/S0140-6736(04)17670-8]

75 **Grimm W**, Müller HH. A randomized controlled trial of the effect of fluid extract of Echinacea purpurea on the incidence and severity of colds and respiratory infections. *Am J Med* 1999; **106**: 138-143 [PMID: 10230741 DOI: 10.1016/S0002-9343(98)00406-9]

76 **Sreedevi A**, Gopalakrishnan UA, Karimassery Ramaiyer S, Kamalamma L. A Randomized controlled trial of the effect of yoga and peer support on glycaemic outcomes in women with type 2 diabetes mellitus: a feasibility study. *BMC Complement Altern Med* 2017; **17**: 100 [PMID: 28173786 DOI: 10.1186/s12906-017-1574-x]

77 **Treasure T**, MacRae KD. Minimisation: the platinum standard for trials?. Randomisation doesn't guarantee similarity of groups; minimisation does. *BMJ* 1998; **317**: 362-363 [PMID: 9694748 DOI: 10.1136/bmj.317.7155.362]

Published by **Baishideng Publishing Group Inc**
7041 Koll Center Parkway, Suite 160, Pleasanton, CA 94566, USA
Telephone: +1-925-3991568
E-mail: office@baishideng.com
Help Desk: https://www.f6publishing.com/helpdesk
https://www.wjgnet.com