RESEARCH ARTICLE

Applied usage and performance of statistical matching in bibliometrics: The comparison of milestone and regular papers with multiple measurements of disruptiveness as an empirical example

Felix Bittmann[1], **Alexander Tekles**[2,3], **and Lutz Bornmann**[2]

[1]Leibniz Institute for Educational Trajectories (Lifbi), Wilhelmsplatz 3, 96047 Bamberg, Germany
[2]Science Policy and Strategy Department, Administrative Headquarters of the Max Planck Society, Hofgartenstr. 8, 80539 Munich, Germany
[3]Ludwig-Maximilians-Universität Munich, Department of Sociology, Konradstr. 6, 80801 Munich, Germany

Keywords: bibliometrics, convergent validity, disruption index, Stata, statistical matching

ABSTRACT

Controlling for confounding factors is one of the central aspects of quantitative research. Although methods such as linear regression models are common, their results can be misleading under certain conditions. We demonstrate how statistical matching can be utilized as an alternative that enables the inspection of post-matching balancing. This contribution serves as an empirical demonstration of matching in bibliometrics and discusses the advantages and potential pitfalls. We propose matching as an easy-to-use approach in bibliometrics to estimate effects and remove bias. To exemplify matching, we use data about papers published in *Physical Review E* and a selection classified as milestone papers. We analyze whether milestone papers score higher in terms of a proposed class of indicators for measuring disruptiveness than nonmilestone papers. We consider disruption indicators DI1, DI5, DI1n, DI5n, and DEP and test which of the disruption indicators performs best, based on the assumption that milestone papers should have higher disruption indicator values than nonmilestone papers. Four matching algorithms (propensity score matching (PSM), coarsened exact matching (CEM), entropy balancing (EB), and inverse probability weighting (IPTW)) are compared. We find that CEM and EB perform best regarding covariate balancing and DI5 and DEP performing well to evaluate disruptiveness of published papers.

1. INTRODUCTION

Scientometric research is mainly empirical research. Large-scale databases (e.g., Web of Science, Clarivate Analytics, or Scopus (Elsevier)) are used to investigate various phenomena in science. An overview of these studies can be found in Fortunato, Bergstrom et al. (2018). A popular topic of scientometric studies is the effect of gender. Researchers are interested in whether gender has an effect on the number of instances of being cited or the chance of being appointed for a professorship or fellowship. They want to know whether there is a systematic and robust gender bias in typical activities in science. Another popular topic of scientometric studies is the effect of the journal impact factor (a journal metric reflecting the reputation of a

an open access journal

Citation: Bittmann, F., Tekles, A., & Bornmann, L. (2021). Applied usage and performance of statistical matching in bibliometrics: The comparison of milestone and regular papers with multiple measurements of disruptiveness as an empirical example. *Quantitative Science Studies*, *2*(4), 1246–1270. https://doi.org/10.1162/qss_a_00158

DOI: https://doi.org/10.1162/qss_a_00158

Peer Review: https://publons.com/publon/10.1162/qss_a_00158

Supporting Information: https://doi.org/10.1162/qss_a_00158

Received: 24 September 2020
Accepted: 21 June 2021

Corresponding Author:
Felix Bittmann
felix.bittmann@lifbi.de

Handling Editor:
Ludo Waltman

The MIT Press

journal) on the citations of the papers published in a journal. Do papers profit from publication in a reputable journal in terms of being cited or not? An overview of studies that have investigated the relationship of journal impact factor and citations can be found in Onodera and Yoshikane (2015). Many of the studies investigating gender bias, citation advantages of the journal impact factor, and other phenomena have used multiple regression models to statistically analyze the data. In these models, the relationships between exactly one dependent variable (e.g., citation counts) and one or multiple independent variable(s) (e.g., journal impact factor) are investigated. Although in general regression methods are a valid tool to estimate (causal) effects, other methods can perform better in certain situations for multiple reasons, which will be outlined further below. In this paper, we present alternative methods—so-called matching techniques—which can be used instead of or as a supplement to regression models. It is our intention to explain the techniques based on a concrete empirical example for possible use in future scientometric studies.

Scientometric data are, as a rule, observational data (and not experimental data). Whenever observational data are available, simply comparing group means can create misleading results due to confounding influences. To achieve unbiased estimations of effects, various matching techniques exist to account for confounding. These techniques are usually referred to as *controlling* or *adjusting* to estimate unbiased effects balancing the distribution of covariates (possibly confounding factors) in the treatment and control groups (Rosenbaum, 1999, 2002; Rubin, 2007). Treatment groups are, for instance, female researchers/papers published by female researchers or papers published in reputable journals. Although statistical matching is not generally superior to methods such as regression models, and results can still be biased, if relevant confounders are omitted, they have several interesting properties that might be able to explain the growing popularity of matching techniques in various disciplines in recent years. These properties are outlined in detail in this study.

A few earlier studies by Farys and Wolbring (2017), Ginther and Heggeness (2020), Mutz and Daniel (2012), and Mutz, Wolbring, and Daniel (2017) have demonstrated how useful matching techniques are for scientometric studies. For example, Mutz et al. (2017) have used the technique to investigate the effect of assigning the label "very important paper" to papers published in the journal *Angewandte Chemie—International Edition*. The authors were interested in whether this assignment has a causal effect on the citation impact of the papers: Do these papers receive significantly more citations than comparable papers without this label? The results show that this is the case. In this study, we build upon these few previous studies and examine various matching techniques. Using a data set from bibliometrics as an exemplary case study, we explain various matching techniques in detail: propensity score-matching (PSM), inverse probability weighting (IPTW), coarsened-exact-matching (CEM), and entropy balancing (EB). The current paper can thus be understood as a methods paper explaining a certain statistic. In our opinion, the scientometric field would profit by applying these techniques more frequently in empirical research.

The example data that we used in this study are from *Physical Review E*—a journal focusing on collective phenomena of many-body systems. Editors of the journal denoted some papers from the journal as milestone papers in 2015[1]. These milestone papers represent the treatment group in the current study. We are interested in whether this group of papers differs from a control group of papers in terms of indicators measuring disruptiveness of research. The goal of our analyses is to test how well the indicators perform: If the indicators adequately identify disruptive papers, the treatment and control group should differ with regard to the indicators.

[1] PRE Milestones (n.d.)

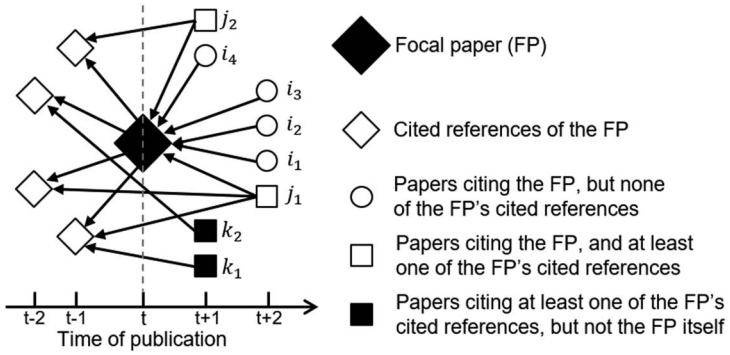
To compare the treatment group with a control group, four matching techniques are applied whereby several confounding variables are controlled in the statistical analyses, such as the number of coauthors of a paper and its number of cited references. The disruption indicators are recent developments in the field of scientometrics. By using the example data set with milestone papers from *Physical Review E,* the current study is a follow-up study of the study by Bornmann and Tekles (2020), who investigated milestone papers of the journal *Physical Review Letters* with the same set of indicators.

In the following sections, the example data set used in this study and the theoretical foundations of matching algorithms are described. Then, the matching results and results of the balancing and robustness checks are reported in the results section. In the last two sections of this paper, the matching procedures are finally discussed in the context of their application for bibliometric studies.

2. DATA SET

For the generation of our data set, we started with a list of milestone papers published in *Physical Review E*. For this list, papers that made significant contributions to their field were selected by the editors of the journal. We assume that papers that made significant contributions to their field are more likely to be the origin of new lines of research (i.e., to be disruptive) than other papers. Based on this assumption, we want to test how well different indicators for measuring disruptiveness perform. To perform well, an indicator should on average differ between milestone and nonmilestone papers (we use the milestone assignment as a proxy for identifying papers that made significant contributions to their field). The papers in the list of *Physical Review E* milestone papers were published between 1993 and 2004. As this list was published in 2015, the selection of milestone papers may be influenced by citation information that was available by then. This possibility must be borne in mind when interpreting the results of our empirical analyses. To complete our data set for this study, we added all papers that are not in the list of milestone papers, but were also published in *Physical Review E* within the same time span. For all these papers, we retrieved additional bibliometric information from an in-house database at the Max Planck Society which is based on the Web of Science. For our analyses, we restricted the data set to the document type "article." This results in a list of 21,164 papers, of which 21 are milestone papers. Hence, the data set is very unbalanced with regard to the classification as milestone paper. Such data sets with a large difference in cases between treatment and control group are rather typical setups for the application of matching techniques. In clinical studies, for example, only a restricted number of ill or treated patients are available, with a large number of potential controls. These kinds of data sets are ideal for matching because the techniques make it possible to select the most appropriate controls out of a large pool of potential controls/donors. As others have pointed out, the control group should be larger than the treatment group by a factor of at least three, as this typically increases the common support region (in PSM) and allows finding multiple controls per treatment case (Olmos & Govindasamy, 2015, p. 86).

As we are interested in the difference between milestone and nonmilestone papers in terms of the indicators measuring disruptiveness, we used these indicators as outcome variables in our study. We considered five different indicators to measure the papers' disruptiveness: DI1, DI5, DI1n, DI5n, and the inverse DEP. These indicators all follow the same idea to measure disruptiveness: A focal paper (FP) can be regarded as disruptive if it is cited by many other papers that do *not* cite the FP's cited references. If this is the case, the citing papers depend on the FP but not its cited references (i.e., one can assume that the FP is the origin of new lines of research). In contrast, papers citing both the FP and its cited references indicate a



$$DI_l = \frac{N_i - N_j^l}{N_i + N_j^l + N_k} \qquad DEP = \frac{N_{j\times cited}}{N_i + N_j^1}$$

$$N_i = |\{i_1, i_2, \ldots\}| \quad N_j^l = |\{j_m | j_m \text{ cites FP and at least } l \text{ of FP's cited references}\}|$$

$$N_k = |\{k_1, k_2, \ldots\}| \quad N_{j\times cited} = |\{(j_m, p_n) | p_n \text{ is cited by } j_m \text{ and FP}\}|$$

Figure 1. Definitions for disruption indexes DI1 and DI5 as well as the dependency indicator (DEP).

developmental FP. This idea of measuring disruptiveness has been introduced recently in the context of patent analysis by Funk and Owen-Smith (2017). Wu, Wang, and Evans (2019) were the first to apply this concept to scientific publications by introducing the indicator DI1. The calculation of DI1 for a given FP is based on three terms (see Figure 1): N_i (the number of papers citing the FP but none of the FP's cited references), N_j^1 (the number of papers citing the FP and at least one of the FP's cited references) and N_k (the number of papers citing at least one of the FP's cited references but not the FP itself). The formula is based on the idea that N_i exceeds N_j^1 if the FP is disruptive. By including N_k, the indicator also considers how strong the citation impact of the FP is compared to its cited references.

Since the introduction of DI1, several modifications of this indicator have been proposed. Out of these modified disruption indicators, we considered DI5, DI1n, and DI5n in this study because they showed good results in existing studies assessing their convergent validity (Bornmann, Devarakonda et al., 2019, 2020; Bornmann & Tekles 2020). In contrast to DI1, DI5 (which was first introduced in Bornmann et al., 2019) considers how strong the ties between the citing and cited side of FPs are: A developmental FP is only indicated by citing papers that also cite at least five (instead of one) of the FP's cited references, which is captured in the term N_j^5 (see Figure 1). DI1n and DI5n are designed to measure the field-specific disruptiveness of a paper (Bornmann et al., 2020). The definitions of DI1n and DI5n correspond to DI1 and DI5, respectively, but the FP's cited references are only considered for determining N_j^5 and N_k if they have been cited by other papers published in the same journal and the same year as the FP. All disruption indicators (DI1, DI5, DI1n, and DI5n) in their original form range from −1 to 1, with high (positive) values indicating disruptive papers (high negative values denote continuity in research). In this study, however, we multiplied the indicators by 100 for the statistical analyses to avoid small numbers with many decimal places. This transformation has been chosen to improve the presentation of the results.

Independently of the development of DI1, DI5, DI1n, and DI5n, Bu, Waltman, and Huang (2021) proposed another indicator (DEP) that also follows the idea of considering whether the citing papers of an FP cite the FP's cited references or not. Like DI5, DEP takes into account how

strong the ties between the citing and the cited side of FPs are. More specifically, DEP is defined as the average number of citation links from a paper citing the FP to the FP's cited references (see Figure 1). A high (average) number of such citation links indicates a high dependency of citing papers on earlier work so that disruptiveness is represented by small values of DEP. In contrast to DI1, DI5, DI1n, and DI5n, DEP does not include a term for assessing the FP's citation impact (relative to the FP's cited references). This corresponds to a different notion of disruptiveness than DI1, DI5, DI1n, and DI5n build upon. DI1, DI5, DI1n, and DI5n follow the idea that FPs need to be relevant for a relatively large set of papers (compared to the FPs' cited references) in order to be disruptive. In contrast, the definition of DEP only considers to which extent citing papers refer to the cited references of FPs. To facilitate the comparison between DEP and DI1, and DI5, DI1n, and DI5n, we use the inverse DEP in this study, which is calculated by subtracting the values of DEP from the maximum value plus 1.

Since the introduction of the disruption indicators, some studies on their behavior and their validity have been published. Bornmann and Tekles (2019) have shown that it may take several years until the values of DI1 for a given paper reach a constant level. Therefore, a sufficiently long citation window is necessary to produce meaningful results (Bornmann and Tekles (2019) suggest a citation window of at least three years). Because the data set of this study only comprises papers that were published in 2004 or earlier, this requirement is fulfilled in our statistical analyses. Other studies have shown that only very few papers score high on DI1, DI5, DI1n, and DI5n, whereas there are usually more papers with high values of the inverse DEP (Bornmann & Tekles, 2020).

Bornmann et al. (2019) examined the convergent validity of the disruption indicators by analyzing the relationship between the indicator values and expert-based tags measuring newness of research. The study by Bornmann and Tekles (2020) used an external criterion for disruptive research similar to the current study to assess the convergent validity of the disruption indicators: a list of milestone papers published in the journal *Physical Review Letters* which were selected by the editors of the journal. Both of these studies found a considerable relationship between the disruption indicators and the external criteria for disruptiveness. However, both studies also found a stronger relationship between the external criteria for disruptiveness and citation impact. A similar finding was reported by Wei, Zhao et al. (2020). The findings of these authors reveal that citation impact is a better predictor for Nobel prize-winning papers than disruptiveness in terms of DI1.

In the current study, we analyze whether milestone papers score higher in terms of the disruption indicators than the other papers published in the same journal. As the milestone papers were selected a few years after their publication, the citation impact may have played a role in the selection process. Therefore, the citation impact is very likely to be a good predictor for milestone papers. At the same time, the definitions of the disruption indicators also depend on citation patterns that may be related to citation impact and variables influencing the citation impact. Thus, citation impact is a confounder for the effect of the milestone variable on the disruption indicators. To focus on this question, we compare the disruption indicator values of milestone and nonmilestone papers, which are comparable aside from the milestone assignment, by controlling the following variables in our analyses. These variables may have a considerable effect on citation impact.

The first variable is the number of coauthors. Due to the effects of self-citations and network effects (Valderas, Bentley et al., 2007), this number might have an effect on citations, as different studies have demonstrated (e.g., Beaver, 2004; Fok & Franses, 2007; Tregenza, 2002; van Wesel, Wyatt, & ten Haaf, 2014) and thus be a potential confounder. In this study, we use

the raw variable with values from 1 to 27. One extreme outlier from the control group with more than 100 coauthors is excluded.

The second control variable is the number of countries involved in a paper, which might have some effects regarding a national citation bias (Gingras & Khelfaoui, 2018). We transform this variable into a binary one (one country versus multiple countries) as there are only very few papers with many countries and it would be difficult to find appropriate matches.

The third variable is the age of each paper in terms of the years since publication. Older papers have had more time to be cited, which might influence their status (Seglen, 1992) and also the disruption indicator score (Bornmann & Tekles, 2019). This variable includes integers ranging from 1 to 12 years since publication.

The fourth control variable is the number of references cited by a paper. Multiple studies have shown a relation between the number of citations and the number of cited references (e.g., Ahlgren, Colliander, & Sjögårde, 2018; Fok & Franses, 2007; Peters & van Raan, 1994; Yu & Yu, 2014). Although presumably not as relevant as in regular regression analyses, we use the log-transformed count of the number of references, as this gives a normally distributed variable which might be beneficial for the CEM cut-off algorithm.

Only papers with complete information on all relevant variables are retained for the statistical analyses (listwise deletion). Because the citation distributions of the milestone papers and the nonmilestone papers in our data set are very different, it is not possible to include the citation impact itself in the matching procedure. By restricting the data set to those papers that have at least as many citations as the least cited milestone paper, it is nevertheless possible to control for citation impact to a certain extent. We additionally used this restricted data set besides the data set including all papers to investigate the robustness of the empirical results.

3. STATISTICAL MATCHING

The general idea behind statistical matching is to simulate an experimental design when only observational data are available to make (causal) inferences. In an experiment, usually two groups are compared: treatment and control. The randomized allocation process in the experiment guarantees that both groups are similar, on average, with respect to observed and unobserved characteristics before the treatment is applied. Matching tries to mimic this process by balancing known covariates in both groups. The balancing creates a statistical comparison where treatment and control are similar, at least with respect to measured covariates. If all relevant confounding factors are accounted for in statistical matching, causal effects can be estimated. Usually, balancing the observed covariates can help to balance unobserved covariates that are correlated with observed ones; hence, balancing is relevant for reaching high quality results (Caliendo & Kopeinig, 2008, p. 18). However, this cannot be proven statistically but must be defended with theoretical arguments. In the following, we present the advantages and challenges of statistical matching. We summarize various techniques that we empirically test using the example data set.

3.1. Advantages and Disadvantages of Statistical Matching

Matching techniques have several advantages (compared to other statistics) for bibliometric analyses:

First, the techniques are conceptually close to the counterfactual framework (Morgan & Winship, 2015): Causal effects are estimated by generating a counterfactual situation whereby cases are observed with the nonfactual status (that is, treatment and control are swapped). In

reality, however, this status does not exist. A case can only either have a treatment status or a control status. Matching approaches nevertheless follow this concept by comparing treated and untreated observations that are comparable with regard to the control variables considered. The idea behind the matching approach is that a treated (untreated) observation would, if it were untreated (treated), behave similarly to an actually untreated (treated) observation with comparable values for the control variables. This means for the empirical example of this study that a milestone paper would behave like a regular paper with similar values for certain control variables (number of coauthors, number of cited references, etc.). The only reason why the two papers behave differently is that one is a milestone paper and the other is not.

Second, the functional form of the relationship between treatment and outcome can be ignored. Although other methods such as linear regressions assume a strictly linear relationship and violations of this assumption can lead to severe biases in the results, matching is agnostic about this relation and reduces the number of specifications that the researcher has to check. This advantage is of special relevance for bibliometrics, as bibliometric data are usually concerned with skewed distributions.

Third, statistical matching allows the user to inspect the quality of the matching, which is an integral aspect of the validity of the estimated effects. Regression models can be considered to be rather opaque, as only regression coefficients are computed. Although the coefficients report the overall effect of a variable under the control of all other independent variables in the model, we are not informed about the validity of the findings. The computed coefficients might be based on highly dissimilar groups, which would invalidate the findings. With matching, the degree of similarity between treatment and control can be assessed after the procedure is performed. It can be examined whether the matching produced highly similar comparison groups or not. If this assumption is violated in matching, the researcher knows that the results must be regarded with uttermost caution (the results probably cannot reveal any unbiased effects). For example, suppose that in a regression model a severe imbalance between treatment and control exists and, even after adjustment, a milestone paper has 10 authors on average and a regular paper only has two. The computed coefficient would be biased because this confounding factor could not be adjusted for. This is invisible to the user, however, who only sees the final coefficient and does not see how the groups were adjusted. Matching designs make these aspects transparent.

Fourth, in comparison to linear regressions that only report a single coefficient, matching allows the computation of multiple estimators with distinct meaning. Average treatment effects (ATEs) correspond to the regression coefficients (betas). ATEs can be interpreted as follows: Suppose a case is randomly selected for treatment. The effect is estimated as the counterfactual effect in comparison to the outcome that would have occurred if the case had been selected for the control group. In other words, the ATE is the effect for the "average" case in the sample. ATEs can be decomposed into ATT (average treatment effect on the treated) and ATC (average treatment effect on the control). ATT is the effect of treatment on those cases that actually received it, and ATC is the counterfactual effect of a case if it would have been treated. Hence, ATE is computed as the weighted mean of ATT and ATC. Depending on the research question, analyzing ATT, ATC, and their difference might be of special interest.

Like all other statistics, matching techniques have several disadvantages that should also be taken into account. The disadvantages are basically the counterparts to the advantages. As neither functional forms nor the separate contribution of control variables can be inspected, these techniques cannot replace regular regression designs. The techniques can be especially used for estimating treatment effects when the concrete functional form between treatment and

outcome is irrelevant. Whenever a treatment is binary, this aspect can be ignored, as there is no functional form to be estimated. For other research questions dealing with *continuous* treatment variables, regression designs might be the better choice. In addition, regression techniques allow for the inspection of effects of multiple independent variables simultaneously, that is, under control of all other independent variables. This makes it possible to estimate how the independent variables *jointly* affect the outcome. In contrast, matching techniques only quantify the effect of the single treatment variable. All other control variables in the model are not further explained or quantified; coefficients are not computed for them. Furthermore, the functional form between treatment and outcome can be estimated using regression models. This functional form can be, for example, linear, quadratic, or exponential, depending on certain assumptions. The selection of the functional form is not possible for matching algorithms; they only compute single treatment effects. However, the functional form is often irrelevant in experimental designs, which matching algorithms attempt to mimic.

3.2. Matching for Causal Inference

Establishing causal relationships is one of the most important yet also most difficult aspects in data analysis, especially for policy-making and evaluation. Matching is a method that facilitates causal inference and especially causality according to Rubin (1974). In our case study, however, we are not interested in the analysis of a (potentially causal) effect of the milestone assignment on disruptiveness. Our goal is to test whether disruption indicators work as they are supposed to work. If this is the case, milestone assignments (a proxy for disruptiveness) should be associated with disruption indicator values. Therefore, matching approaches are a reasonable choice in this situation, because they allow us to control for the possible confounders mentioned in Section 2. By controlling confounders, associations between milestone assignments and disruption indicator values would not be due to confounding of control variables. Using matching approaches also allows us to assess matching quality. This is important in our case given the large control group (see also the advantages of matching approaches mentioned in Section 3.1).

With regard to using matching approaches for causal inference, we encourage the reader to have a look at the steadily growing body of literature and especially consult the works of Imbens and Rubin (2015), Morgan and Winship (2015), Pearl (2009), and Pearl, Glymour, and Jewell (2016). The authors target the social sciences and provide detailed examples. A nontechnical introduction for laypersons is given by Pearl and Mackenzie (2019). Whether or not the results of matching can be interpreted as causal effects depends on whether researchers are able to establish thoroughly that all assumptions for causal inference are indeed fulfilled. This can be achieved by theoretical and careful argumentation: No statistical test can derive whether or not a result is a causal effect. When researchers are not able to argue convincingly that all requirements are met, they should highlight the associational character of the findings. They cannot rule out hidden variable bias (for example).

3.3. An Overview of Various Matching Algorithms

After explaining the advantages and disadvantages of matching techniques in general, we present in the following an overview of various matching algorithms and explain their approaches to generate a balanced sample. Depending on the research questions, data sets, and designs of a certain bibliometric study, one of the matching algorithms might yield the most robust results. In this study, we apply four algorithms to the example data set; however, this is usually not feasible in a typical bibliometric paper. Our suggestion is therefore to

compare at least a few algorithms with quite different statistical approaches (for example, CEM and EB) and inspect the quality of the findings. The selection of the algorithm should then be based on the most stable findings.

3.3.1. Propensity score matching (PSM)

To model the selection into treatment, a logistic (alternatively probit) model is used where the binary treatment status is the dependent variable and all potential confounders are independent variables. The model computes the individual probability for each case to be selected for treatment as a number between 0 and 1 (Rosenbaum & Rubin, 1983). Because the potential confounders are relevant for the score, a case with a high individual propensity score has a high probability of being selected for treatment, even if the factual status is the control condition. Before matching, the region of common support for both treatment and control group should be reviewed: the computed propensity scores are compared between the groups. Only those cases are retained that have a value that is also available in the other group. For example, when the propensity score ranges from 5 to 60 in the control group and from 10 to 75 in the treatment group, the region of common support is from 10 to 60. There are no clear guidelines in the literature about whether imposing this restriction is always necessary, as it usually leads to a reduction of available cases. Modern implementations, in particular, of PSM, such as kernel-matching, usually do not benefit much from this restriction. In the analyses of this study we impose the common support restriction. After computing and restricting the propensity scores, cases are matched on it. For each case in the treatment group, one or multiple cases from the control group are selected, which should have an identical or very similar score.

Nearest-neighbor matching selects up to *n* neighbors for each treated case (Caliendo & Kopeinig, 2008). It is probably the most popular derivation of the general matching idea, as the assumptions are easy to comprehend, and it is implemented in many statistical software packages. By introducing a caliper (the maximum distance of two neighbors with respect to the propensity score), results can be improved as bad matches are avoided. By setting the caliper the user can adjust the balance between finding many matches and finding especially close matches. The mean differences in the outcome variable between matched cases can be compared to estimate the unbiased effect of the treatment. A similar propensity score guarantees that, on average, the cases are similar with regard to all control variables. More recent implementations rely on kernel instead of nearest-neighbor matching. Here, instead of selecting *n* neighbors, every single case is used but weighted by the degree of similarity (Bittmann, 2019). The closer the propensity score of a neighbor, the larger the weight. Although the introduction of kernel weighting usually improves the performance, reported case numbers can be deceptively large when many cases receive a weight close to zero (and contribute basically nothing to the estimation). Let us explain the technique based on our example data set. Instead of finding some similar control papers for a milestone paper which should be the nearest neighbors with respect to the propensity score, every single control paper is utilized as a neighbor. Then, only those control papers with a similar propensity score receive a high weighting, and other control papers with a highly different propensity score are discounted and receive a lower weighting. A very early implementation of the PSM approach is described in Rosenbaum and Rubin (1985). Further basic information on the approach can be found in Abadie and Imbens (2016), Heinrich, Maffioli, and Vazquez (2010), and Morgan and Winship (2015). If subgroup analyses are of interest in a study, these should be matched separately.

For the practical application of the technique, various software programs are available such as SPSS (Thoemmes, 2012), Stata (Jann, 2017a), and R (Randolph, Falbe et al., 2014; Olmos & Govindasamy, 2015). Although nowadays PSM is probably the most popular among the

matching algorithms, some researchers argue that it might lead to an *increased* imbalance between groups (King & Nielsen, 2019) and might be inefficient (Frölich, 2007). Others counter that these downsides are only valid for rather crude PSM variants (one-to-one matching without replacement) and more recent implementations such as kernel matching do not display these problems (Jann, 2017b). In any case, due to its overall popularity and widespread use, we include PSM in this study and compare its performance with other algorithms. A further option to consider is the usage of regression adjustment, that is using the computed propensity score as a further control variable or stratifying the analyses based on propensity score levels (D'Agostino Jr., 1998).

3.3.2. Inverse probability weighting (IPTW)

Similar to PSM, IPTW relies on the propensity scores, which are calculated as described above; the same rules hold for selecting a region of common support. Each case receives a weight which is the inverse of the probability of receiving the factual status (Horvitz & Thompson, 1952). For example, case n_i in the treatment group receives the weight $w_i = 1/p_i(\text{Treatment})$, whereby $p_i(\text{Treatment})$ is the individual propensity score of this case. Cases in the control group receive the weighting $w_i = 1/[1 - p_i(\text{Treatment})]$. That means that a case with a low probability of treatment in the treatment group receives a high weighting because it is similar to the untreated cases and enables a comparison. Cases with a high probability of treatment in the treatment group are weighted down, as there are many similar cases available with the same status. The calculation of the effect is then the weighted difference of means between the two groups. More information on the technique can be found in Austin and Stuart (2015) and Halpern (2014).

3.3.3. Coarsened exact matching (CEM)

Instead of relying on a propensity score, CEM attempts to find perfect matches. A perfect match occurs when there is a case available with a different treatment status but otherwise exactly the same characteristics (e.g., the same number of coauthors). Because the "curse of dimensionality" usually prevents the finding of perfect matches when the number of control variables is large, coarsening is used as a potential remedy (Iacus, King, & Porro, 2012). For example, a continuous variable with a large number of distinct values is coarsened into a prespecified number of categories, such as quintiles. Matching is then performed based on quintile categories and the original information is retained. After matching based on the coarsened variables, the final effects are calculated as differences in the outcome variable between group means using the original and unchanged dependent variable.

The finer the degree of coarsening, the lower the number of potential matches. It is up to the user of CEM to test different coarsening rules and to find a balance between large numbers of matches and high levels of detail and matching precision. For creating and selecting categories, multiple rules and algorithms are available. Suppose, for example, a user matches treatment and control papers based on their citation counts. As citation counts is a continuous variable, it might be impossible to find a perfect match for a paper with a specific number of citations, because no other paper in the control group has exactly this number. However, another paper is available having just one citation more. Through coarsening based on quintiles, both papers end up in the same quintile (a group of papers within a certain range of citation counts). The treatment paper with the specific number of citations has a match, therefore—albeit not a perfect match.

By coarsening, the aforementioned "curse of dimensionality" can be greatly ameliorated when many independent variables are included in a model. In our example data set, the

binary variable "number of countries" is matched perfectly (because there are only two categories available and further coarsening is impossible). For more information on how to apply CEM, including practical examples, see Guarcello, Levine et al. (2017), Schurer, Alspach et al. (2016), and Stevens, King, and Shibuya (2010).

3.3.4. Entropy balancing (EB)

In contrast to PSM, IPTW, and CEM, EB turns around the matching process. Instead of selecting similar cases and testing for balance afterwards, EB forces balancing with respect to prespecified conditions and generates matches according to the constraints by reweighting cases (Hainmueller, 2012). As this technique is highly flexible, the user can select various statistical moments that must be matched. These moments are usually means (first moment) and variances (second moment) of the independent variables. EB can be generalized to higher moments as well and some statistical packages allow matching of the skewness or even covariances.

After selecting the constraints, a loss function is used to meet the constraints. Each case receives a weight that is applied when group differences are computed. Constraints might not be met due to small sample sizes, a large number of constraints (matching multiple moments and covariances), or a strong imbalance between treatment and control group. If the constraints are not met, the algorithm does not converge and cannot yield an estimation. As a possible solution, the user can reduce the number of constraints. If the algorithm converges, the specified moments are basically guaranteed to be equal. The balancing should be close to an ideal state. A failure of balancing here might be a good indication for the user that other matching methods also provide suboptimal results. Further information on EB is available in Abadie, Diamond, and Hainmueller (2010), Amusa, Zewotir, and North (2019), and Zhao and Percival (2016).

3.4. Software

All the results presented in the following are computed using Stata 16.1 and the user-written software package *kmatch* (Jann, 2017a), which implements all of the matching algorithms described above. In the supplemental material, we also provide results computed using R as an additional robustness check (and to demonstrate that R can be equally used for matching as Stata). For the R analyses, we used the R packages *MatchIt* (Ho, Imai et al., 2011), *ebal* (Hainmueller, 2014), and *boot* (Canty & Ripley, 2021).

4. RESULTS

4.1. Descriptive Statistics

Table 1 presents basic descriptive statistics for the milestone and regular papers included in this study. Although the asymmetry regarding the number of milestone papers to regular papers is extreme, the distribution of the control variables is very similar. For example, the number of coauthors involved and the number of cited references is comparable and not statistically significantly different between milestone and regular papers. Only the time since publication is statistically significantly different between both groups. In contrast to most of the control variables, most outcome variables display statistically significant differences between regular and milestone papers.

Figure 3 presents distributions of the outcome variables graphically using histograms. The histograms show that most of the values for DI1, DI5, DI1n, and DI5n lie in a small range

Table 1. Descriptive statistics for the entire sample

	Minimum	Maximum	Mean	Standard deviation	Median
			Milestone papers (N = 21)		
Multiple countries involved	0.000	1.000	0.381	0.498	0.000
Number of coauthors	1.000	6.000	2.905	1.480	3.000
Years since publication (2005)	1.000	12.000	7.048*	3.263	7.000
Logarithmized number of cited references	2.565	4.331	3.516**	0.486	3.497
DI1 (DV)	−10.306	27.217	0.953	9.888	−2.826
DI5 (DV)	−0.663	32.702	7.333***	11.291	1.893
DI1n (DV)	−0.072	0.085	−0.023***	0.037	−0.030
DI5n (DV)	−0.028	0.125	0.015***	0.038	−0.001
DEP (inverse) (DV)	28.176	30.742	29.779**	0.815	29.962
			Regular papers (N = 21,143)		
Multiple countries involved	0.000	1.000	0.468	0.499	0.000
Number of coauthors	1.000	27.000	2.815	1.644	2.000
Years since publication (2005)	1.000	12.000	5.593	3.215	5.000
Logarithmized number of cited references	0.000	5.094	3.217	0.506	3.219
DI1 (DV)	−64.516	91.566	−0.636	2.676	−0.313
DI5 (DV)	−15.385	93.902	0.453	2.884	0.000
DI1n (DV)	−0.115	0.102	−0.001	0.003	−0.001
DI5n (DV)	−0.024	0.215	−0.000	0.002	−0.000
DEP (inverse) (DV)	1.000	31.000	28.325	2.127	28.765

Notes. Asterisks in column "Mean" indicate whether group differences between regular and milestone papers are statistically significant (based on *t*-tests).

*$p < .05$, **$p < .01$, ***$p < .001$

Variables that are used as dependent variables in this study are marked DV.

around 0. There are only a few papers with relatively large or small values for these indicators. In contrast, the distribution for the inverse DEP indicator is less concentrated, even though most papers have values greater than 20. These results are in accord with the results of other empirical analyses concerned with disruption indicators (Bornmann & Tekles, 2020).

In addition, we use kernel-density plots to visualize how the citation counts differ between regular and milestone papers (see Figure 2).

The results in the figure reveal that milestone papers are cited more frequently than regular papers. Because we cannot include citation counts as a further control variable (see above), we run robustness checks where we remove all regular papers that have logarithmized citation

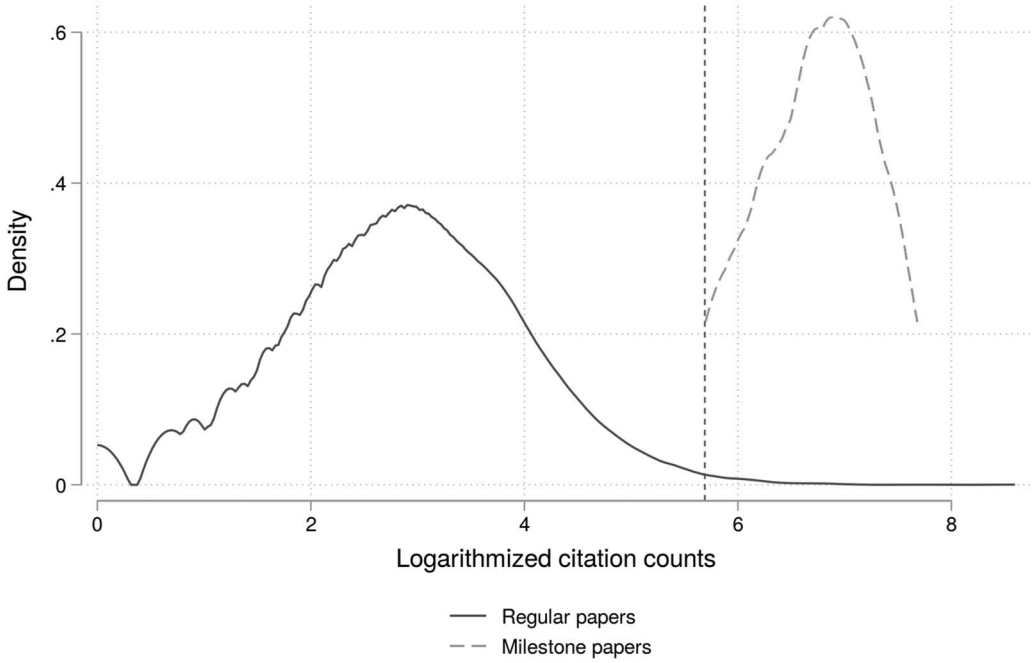


Figure 2. Comparison of regular and milestone papers with respect to logarithmized citation counts. The lower limit of milestone papers is indicated by the vertical bar.

counts below the lowest value of a milestone paper (5.69), which we indicate in the figure using a vertical bar[2].

4.2. Balancing and Number of Cases Used

Before we discuss the treatment effect estimates of the various matching techniques based on the example data set in the following sections, we inspect the balancing, as this is relevant for judging the quality of the findings. All matching algorithms make it possible to inspect how well the observed covariates are balanced between treatment and control group. This is done by applying the computed weight to each case and recalculating the summary statistics of all independent variables. Balancing all control variables is a relevant aspect to obtain valid results. Even with balanced covariates, however, unmeasured variables might still be unbalanced and affect the validity of the estimation. When the balance of other influences (variables) is not approximated, a "fair" comparison between the groups is not possible as pretreatment differences are not completely accounted for. For a convenient interpretation of the balancing results, we create a single figure including all relevant information. We check the balancing for means, variances, and skewness. The means are the most relevant outcomes, as they are the first moment and determine the general shape of a distribution. The results are depicted in Figure 4.

The covariates are enumerated from 1 to 4 (1 = number of countries, 2 = number of coauthors, 3 = number of cited references, 4 = number of years since publication). "Means" reports the standardized difference between milestone and regular papers. When we look at the results for the PSM algorithm, we notice that differences regarding the means are quite large

[2] Our initial analyses have shown that none of the matching algorithms is able to find an acceptable number of matches when this variable is included as independent variable. Therefore, we decided to use this approach.

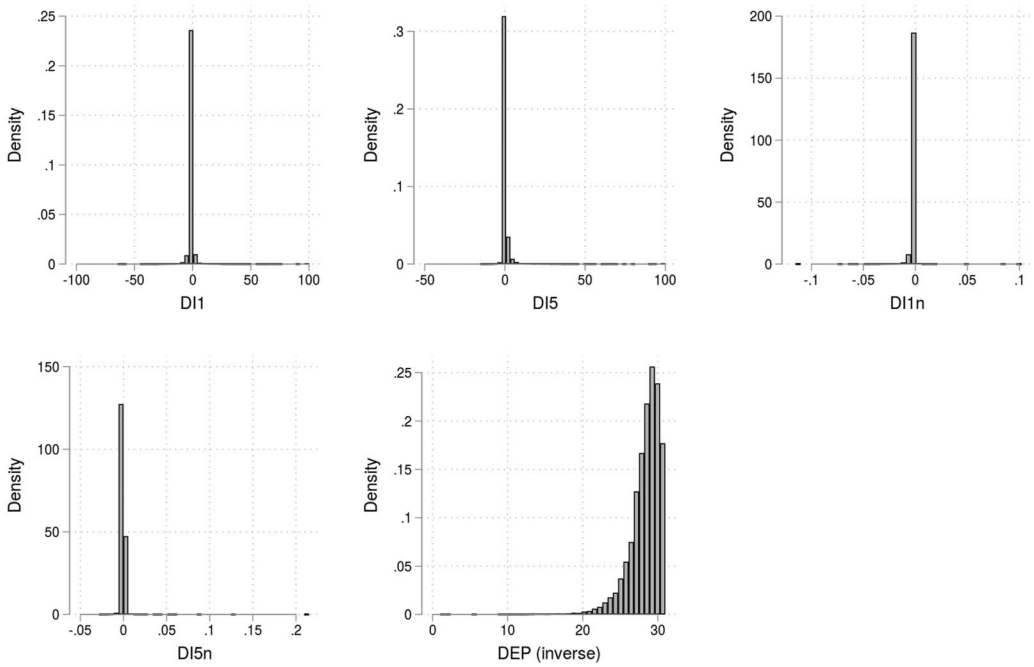


Figure 3. Distributions of all dependent variables.

and can go up to 0.3. A perfect result would be close to zero. For the variances, the deviations are smaller for most variables (a result of 1 would be ideal because we look at the ratios for this variable). For the third moment, the skewness, a few differences are large. We conclude that even after running the PSM models, some differences between the treatment and control groups remain. Perfect comparability with respect to all independent variables in the models cannot be guaranteed.

4.3. Results of the Matching Techniques

The actual matching outcomes of the four techniques are presented in Table 2. The table reports the average treatment effect (ATE) to measure the overall effect of treatment. Standard errors are computed analytically using influence functions (Jann, 2019). To test robustness, 95% confidence intervals are provided for the ATEs using bias-corrected bootstrapping with 2,000 resamples (Bittmann, 2021; Efron & Tibshirani, 1994). Because analytical standard errors can be too conservative for matching, we can test whether the conclusions are the same for both forms of computation (analytical and bootstrap standard errors) (Austin & Cafri, 2020; Hill, 2008; Jann, 2019). With ATEs, it is usually not possible to compute pure causal effects when only observational data are at hand and not all potential confounders are available. Therefore, the findings below can only be interpreted as rather associational than causal. To enable a comparison with the popular regression-based approaches, we also provide estimates for the treatment effects using ordinary least squares regression models in the supplemental material (Table S1).

Table 2 also reports the number of cases used in the statistical analysis. We notice that only CEM actually prunes many cases with a bad match (lower number of cases used). This is the only technique actually discounting controls, which are quite dissimilar with respect to the characteristics of their independent variables. All other techniques rely on some form of weighting and bad matches receive a very low weight. This means—as Table 2 reveals—that the estimated

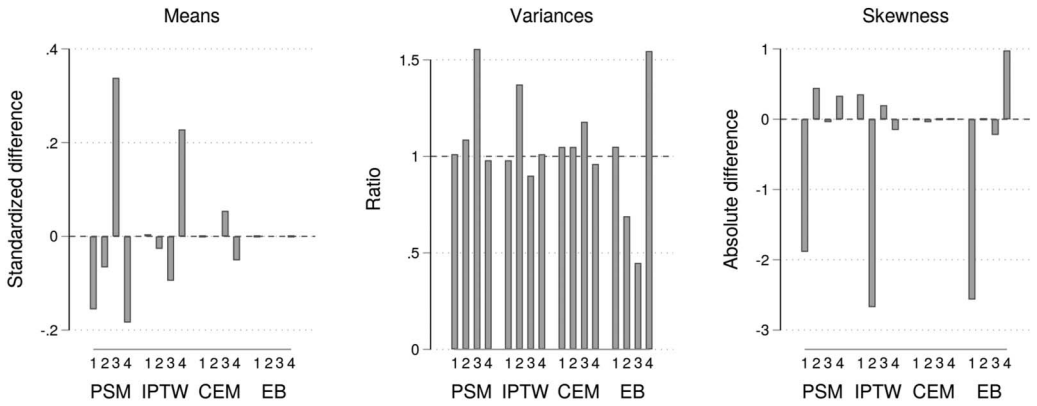


Figure 4. Inspecting balancing with respect to all independent variables between regular and milestone papers. The IVs are enumerated from 1 to 4 (1 = number of countries, 2 = number of coauthors, 3 = number of cited references, 4 = number of years since publication).

relationship between the milestone assignments and the indicator values is only based on very few papers with particular characteristics.

For the robustness check of our results (see Table 3), we compute the same models (including all matching techniques) but exclude all cases from the control group with rather low citation counts to enable a fair comparison (see above). This selection process drastically reduces the case numbers.

4.3.1. Propensity score-matching (PSM)

We utilize a logistic model to compute the propensity score and kernel-matching to estimate ATEs. We restrict the region of common support and give a graphical representation of this process in the supplemental material (see Figure S1). This procedure is identical for PSM and IPTW. A review of the most popular software packages in Stata and R reveals that restricting the common support when applying kernel matching is not used as default and should only be imposed by the researcher if necessary.

The results in Table 2 show that PSM loses some cases due to restricting the common support region. The results reveal that five indicators have a statistically significant result when regular standard errors are computed: DI5, D15n, DI1n, the inverse DEP, and the logarithmized number of citations. The negative ATE of DI1n is an unexpected result, but probably not substantial. The bootstrap confidence interval (CI) does not agree, as zero is included in the interval. We cannot conclude, therefore, that a true relation is present. For the other statistically significantly independent variables the results of both CIs agree. According to the PSM technique, these three indicators should be rather robust. To test the stability of these findings, one option is to compute Rosenbaum bounds as a sensitivity analysis (Rosenbaum, 2002). The basic principle is to simulate the effect of unobserved variables on the selection into treatment and how this affects the results. If even a small additional effect of potential unobserved factors invalidates the findings (by changing *p*-values drastically), the results are probably not stable. We provide an exemplary analysis for the outcome variable DI1n in the supplemental material, see Table S2. When Gamma is 1, the *p*-value approximates the *p*-value of the average treatment effect from PSM as no unobserved influence is specified. The critical value of 0.05 is reached with a Gamma of 1.7: a change of 0.7 in odds in treatment assignment produces a statistically different result than the observed one. The larger the critical Gamma, the more robust the findings are with respect to unobserved influences that affect the treatment status (DiPrete and Gangl, 2004). As the value of 1.7 is not close to 1, we assume that the

results are stable with respect to unobserved influences, even when the bootstrap CI is inconclusive.

The robustness check of the results (see Table 3) shows that only the citation count keeps its statistically significant result and DI5n is very close. Based on the results in Table 2 and their robustness checks in Table 3, we conclude that there are at most two disruption indicators with statistically significant results using the PSM technique. It should be considered in the interpretation of the results, however, that the balancing is not optimal.

Table 2. Matching results

	PSM	IPTW	CEM	EB
DI1				
ATE	1.5786	1.4077	3.0627	1.9628
SE	(2.3082)	(1.9411)	(2.9518)	(2.1160)
95% analytical-CI	[−2.945; 6.102]	[−2.397; 5.212]	[−2.723; 8.848]	[−2.185; 6.110]
95% bootstrap-CI	[−2.787; 7.160]	[−2.762; 5.697]	[−0.861; 7.346]	–
DI5				
ATE	7.5612**	7.0657**	7.9726*	6.8608*
SE	(2.8426)	(2.5656)	(3.2481)	(3.2695)
95% analytical-CI	[1.989; 13.132]	[2.036; 12.094]	[1.606; 14.339]	[0.452; 13.269]
95% bootstrap-CI	[2.465; 14.362]	[2.585; 12.734]	[3.570; 13.159]	–
DI1n				
ATE	−0.0198**	−0.0197**	−0.0148	−0.0159*
SE	(0.0070)	(0.0065)	(0.0119)	(0.0069)
95% analytical-CI	[−0.033; −0.006]	[−0.032; −0.0069]	[−0.038; 0.009]	[−0.029; −0.002]
95% bootstrap-CI	[−0.033; 0.002]	[−0.0328; −0.0058]	[−0.030; 0.003]	–
DI5n				
ATE	0.0148*	0.0144*	0.0209	0.0135
SE	(0.0068)	(0.0068)	(0.0123)	(0.0077)
95% analytical-CI	[0.001; 0.028]	[0.001; 0.027]	[−0.003; 0.045]	[−0.001; 0.029]
95% bootstrap-CI	[0.002; 0.035]	[0.003; 0.030]	[0.006; 0.038]	–
DEP (inverse)				
ATE	1.7955***	1.7756***	1.7352***	1.6957***
SE	(0.1614)	(0.1509)	(0.2127)	(0.1987)
95% analytical-CI	[1.479; 2.111]	[1.479; 2.071]	[1.318; 2.152]	[1.306; 2.085]
95% bootstrap-CI	[1.315; 2.098]	[1.385; 2.204]	[1.437; 2.111]	–

Table 2. (*continued*)

	PSM	IPTW	CEM	EB
Logarithmized citation counts				
ATE	3.7225***	3.6804***	3.7807***	3.6061***
SE	(0.1502)	(0.1276)	(0.1289)	(0.1339)
95% analytical-CI	[3.427; 4.016]	[3.430; 3.930]	[3.528; 4.033]	[3.345; 3.868]
95% bootstrap-CI	[3.464; 4.017]	[3.443; 3.952]	[3.537; 3.958]	–
N match (treated)	21	21	21	21
N match (control)	16,947	17,465	990	21,143

Notes. CI = Confidence interval, ATE = Average treatment effect, SE = Standard error, PSM = Propensity score matching, CEM = Coarsened exact matching, EB = Entropy balancing, IPTW = Inverse probability weighting. The outcome variables are the various disruption indicators (and citation counts), which can be found in the column on the left side.

*$p < .05$, **$p < .01$, ***$p < .001$

4.3.2. *Inverse probability weighting (IPTW)*

With respect to the main findings in Table 2, we see that DI5, DI1n, DI5n, the inverse DEP, and citation counts display statistically significant coefficients. The table also shows a stable *negative* association for DI1n which is against our expectations. This means that milestone papers have a lower DI1n than the papers in the control group (on average). When we take a look at the robustness checks in Table 3, this significance vanishes and only the findings of DI5 and citation counts remain stable.

The quality of the balancing indicates that the deviations for IPTW are the second largest after those for PSM. This result concerns the means but also the other moments. The balancing does not seem to be optimal even after the matching was performed. Thus, it appears that the IPTW results are not trustworthy.

4.3.3. *Coarsened-exact-matching (CEM)*

For control variables that are considered to be continuous, Doane's algorithm (1976) is selected to create categories. This formula does not simply generate equally spaced bins, but classifies cases into categories based on the distribution of the variable. Many other algorithms besides Doane's algorithm exist for this purpose. As no general standard has emerged hitherto, it is up to the user to try to compare various options for optimal results. We test different operationalizations and decide in favor of Doanes's formula because the balancing of means and variances gives the best results.

The results in Table 2 show that DI5, the inverse DEP, and citation counts are statistically significant. For these variables, regular and bootstrap CIs agree, highlighting the stability of the results. The inspection of the robustness checks (see Table 3) reveals that DI5 is no longer statistically significant, DI5n becomes statistically significant, and citation counts remain statistically significant. Due to the very low number of cases used in the robustness checks (only nine in total), it is not feasible to compute bootstrap CIs. The regular results for the CEM are stable and robust; the robustness check might be neglected as the case number is very low.

The balancing in Figure 4 indicates that the deviations from the optimal results are very small for CEM. Treatment and control groups are very similar regarding the independent

variables after the matching was performed. These results indicate the high quality of the matching process.

4.3.4. Entropy balancing (EB)

In our example data set, we select only the first moment (arithmetic means) as constraint as the model does not converge when we include higher moments as well. We assume that this is due to the very low number of milestone papers.

Table 3. Matching results (restricted sample)

	PSM	IPTW	CEM	EB
DI1				
ATE	3.5996	3.6531	4.2745	3.8060
SE	(2.3975)	(1.9176)	(6.1216)	(2.0458)
95% analytical-CI	[−1.135; 8.334]	[−0.133; 7.440]	[−7.815; 16.364]	[−0.234; 7.846]
95% bootstrap-CI	[−0.233; 9.488]	[0.118; 8.391]	–	–
DI5				
ATE	4.2480	4.6297*	7.3048	4.5821
SE	(2.4183)	(2.1567)	(6.3376)	(2.3608)
95% analytical-CI	[−0.527; 9.023]	[0.370; 8.888]	[−5.211; 19.821]	[−0.080; 9.244]
95% bootstrap-CI	[−0.095; 10.357]	[0.880; 9.915]	–	–
DI1n				
ATE	−0.0042	−0.0041	0.0222	−0.0018
SE	(0.0086)	(0.0073)	(0.0253)	(0.0076)
95% analytical-CI	[−0.021; 0.012]	[−0.018; 0.010]	[−0.0278; 0.072]	[−0.0169; 0.013]
95% bootstrap-CI	[−0.020; 0.017]	[−0.0179; 0.0136]	–	–
DI5n				
ATE	0.0139	0.0139	0.0526**	0.0127
SE	(0.0082)	(0.0072)	(0.0191)	(0.0078)
95% analytical-CI	[−0.002; 0.030]	[−0.001; 0.028]	[0.015; 0.090]	[−0.003; 0.028]
95% bootstrap-CI	[−0.001; 0.040]	[0.001; 0.032]	–	–
DEP (inverse)				
ATE	0.3430	0.3653	0.5410	0.4506*
SE	(0.1807)	(0.1871)	(0.3963)	(0.1791)
95% analytical-CI	[−0.013; 0.699]	[−0.004; 0.734]	[−0.242; 1.324]	[0.097; 0.804]
95% bootstrap-CI	[−0.2124; 0.689]	[−0.034; 0.704]	–	–

Table 3. (*continued*)

	PSM	IPTW	CEM	EB
Logarithmized citation counts				
ATE	0.5472***	0.5410***	0.7542*	0.4887***
SE	(0.1344)	(0.1262)	(0.3694)	(0.1256)
95% analytical-CI	[0.2817; 0.812]	[0.291; 0.790]	[0.025; 1.484]	[0.241; 0.737]
95% bootstrap-CI	[0.256; 0.841]	[0.296; 0.825]	–	–
N match (treated)	20	21	4	21
N match (control)	131	133	5	140

Notes. CI = Confidence interval, ATE = Average treatment effect, SE = Standard error, PSM = Propensity score matching, CEM = Coarsened exact matching, EB = Entropy balancing, IPTW = Inverse probability weighting. Some confidence bands could not be calculated due to technical reasons. All regular papers with logarithmized citation counts below 5.69 are excluded. The outcome variables are the various disruption indicators (and citation counts), which can be found in the column on the left side.

*$p < .05$, **$p < .01$, ***$p < .001$

Table 2 indicates that the results for DI5, DI1n, the inverse DEP and citation counts are statistically significant. The negative association of DI1n is against the expectation. As the ATE is small and weak, it is probably not a robust finding. The coefficients for the inverse DEP and citation counts are large. The robustness checks in Table 3 show that only the inverse DEP and citation counts keep their statistical significances. The inspection of the balancing in Figure 4 reveals that deviations regarding the means are extremely small. This can be expected because the first moment can be matched very well. The figure also shows that the deviations for the variances and skewness are larger and worse than for CEM.

When we compare the results of EB with the results of CEM, we can conclude the following: Although the means are matched perfectly, the larger deviations with respect to variances and skewness speak for CEM. The balancing is better overall with CEM. Potential biases are probably smaller using CEM (i.e., the resulting conclusions are stronger using this algorithm).

5. DISCUSSION

In this paper, we demonstrate how statistical matching techniques can be utilized as an addition or alternative to other methods, such as linear regressions. In contrast to these other methods, matching techniques are not only closer to the counterfactual framework but can sometimes be more adequate for analyses where the effect (or association in a noncausal framework) of exactly one (binary) treatment variable is of interest. Due to the different statistical approach, researchers not only estimate the desired statistic (in most cases the ATE) but are also able to study in detail how well treatment and control are matched after the procedure is performed. In contrast to linear regressions, where this aspect is opaque, researchers are able to conclude how well the matching was performed for the control variables specified and whether any larger bias is to be expected. By doing so, the quality of the results can be tested which is clearly highly relevant for scientific progress (in scientometrics). It is another advantage of statistical matching that the functional form between treatment and outcome can be ignored. We demonstrate in this study how matching can be applied in a practical example. We utilize bibliometric data to test which disruption indicator performs best. Several control variables are included to account for spurious correlations.

In this study, we use an example data set based on *Physical Review E* papers to demonstrate several matching techniques: PSM, IPTW, CEM, and EB. PSM and IPTW rely on the computation of the propensity score. This score is based on the control variables and predicts the propensity to be in the treatment group. CEM and EB have different requirements than PSM and IPTW. CEM implements an exact matching on broader categories. Depending on how the cut-off points for these categories are chosen, researchers are able to find a balance between precision and the number of cases left for analysis. EB attempts to force the balancing of covariates in advance. The balancing can fail to converge, however, if the number of cases is small and a good balancing solution is not feasible. If this happens, researchers can try to relax the balancing assumptions and can match only means and not variances.

In the empirical case study, we test with the matching techniques whether milestone papers differ from nonmilestone papers with regard to the various disruption indicators. Our results show that DI5, the inverse DEP and logarithmized citation counts have the strongest and robust results whereas outcomes for the other indicators are rather mixed. This suggests that these indicators perform best with regard to measuring the disruptiveness of papers. The found strong association for the number of citations is in line with the results from other studies (Bornmann et al., 2019; Bornmann & Tekles, 2020). These results show that citation impact should ideally be controlled in the matching process to assess whether milestone and nonmilestone papers differ with regard to the disruption indicators. In this study, this is not possible, as the citation distributions of the milestone and nonmilestone papers are very different in our data set.

Because citation counts themselves can not be included as control variable in the matching approaches, we performed robustness checks by restricting the papers to those with citation counts at least as high as the citation counts for the least cited milestone paper. This procedure makes it possible to control for citation impact to a certain degree. Among the disruption indicators, DI5 seems to perform best. This accords with existing studies that also find promising results for DI5 (Bornmann et al., 2019; Bornmann & Tekles, 2020). In contrast to the existing studies, however, we also find promising results for the inverse DEP. The fact that DI5 and the inverse DEP perform best in our analyses may suggest that indicators measuring disruptiveness should take into account how strong the relationships between a citing paper and the cited references of a focal paper are, instead of only considering whether there is a citation link or not. Although DI5 and the inverse DEP both follow this idea, the approach to measure the field-specific disruptiveness of a paper (DI1n and DI5n) does not seem to be useful.

In this paper, we report results computed in R (in addition to Stata results) using the packages *MatchIt* (Ho et al., 2011), *ebal* (Hainmueller, 2014), and *boot* (Canty & Ripley, 2021); see supplemental material Table S3. The additional results reveal that the two software packages come to comparable results. The results underline that the implementations of matching techniques are equivalent and do not influence the conclusions.

In the application of the matching techniques in the empirical analyses, one is usually interested in which algorithm works best with the data. In this study, CEM and EB have the most robust and stable findings overall as well as the smallest deviations when looking at the balancing scores (see Figure 4). Here, deviations for the mean are small for both CEM and EB, which is the most relevant aspect when analyzing balancing statistics. Strong balance for derived statistics such as the variance and the skewness is also preferable but less relevant than balanced means. As both algorithms do rather well for all three measures (small deviations from a perfectly balanced sample after matching), we conclude that these two should be

utilized as they minimize bias. In other words, both algorithms produce the best and most valid findings for our data set. PSM and IPTW display larger deviations regarding the matching of means for most variables. Even after the matching is performed, the difference of covariates between treatment and control group is comparably large. This can lead to biased and wrong conclusions. The propensity score might play a role in this context as this aspect is common to both. It is an advantage of matching techniques that we are able to test balancing and make this crucial aspect of the analyses visible (i.e., we can judge the final quality of the findings). This is not possible with many other techniques, such as linear regressions.

Previous studies compare the results of at least two techniques (Ginther & Heggeness, 2020) or combine different matching techniques (Farys & Wolbring, 2017). We would like to encourage researchers to follow these examples and we suggest applying more than one algorithm and comparing results. Modern software packages make it convenient to compute various algorithms. Researchers strengthen the robustness of their results by doing so. Another option is to combine regression models with propensity score matching, which is referred to as a "double robust estimator" (Funk et al., 2011).

An important research gap that should be tackled in future studies is to compare algorithms systematically and to investigate how they perform with different scientometric data sets (e.g., with respect to the size of the treatment and control groups, total number of cases, and the number and kind of control variables used). Simulation studies might be helpful to find optimal algorithms for their analyses. The idea of such simulations is to generate a data set with known, prespecified effects which are set by the researcher. Then, the matching algorithms are applied to the data set to analyze whether they can recover the baseline truth. By repeating these simulations many times with varying conditions, the strengths and weaknesses of different algorithms can be tested systematically. As the number of potentially conceivable data sets is infinite, one would have to set very clear conditions. These conditions refer to the factors that should be evaluated and the specifications of the performance measurements for the algorithms. An example of using simulated data to validate a certain method in the field of scientometrics can be found in Milojević (2013), who applied this approach to assess different algorithms for identifying authors in bibliometric databases.

What are the limitations of our empirical analyses (the use of matching techniques in scientometrics)? Of course, it is not possible to report unbiased causal effects as only observational data are usually available in scientometrics (another assumption that it is necessary to inspect in a causal framework is strong ignorability). In addition, it is not possible to include every single confounding variable in a study that would be relevant in principle. In this study, for example, the citation impact can not be considered in the matching approaches. This is problematic because the disruption indicator values may be related to the number of citations that a paper receives and at the same time there is a strong relationship between citation impact and milestone paper assignments. Therefore, the estimated ATEs (without considering the citation impact as control variable) can be confounded by the citation impact. To still account for confounding by citation impact, we control for other variables that are related to citation impact. Another limitation concerns the extreme imbalance of the number of papers in the treatment and control groups. For the application of matching techniques, more balanced data sets should be used (ideally).

As in every scientometric study, empirical results must be interpreted with caution. The results of the current study give an estimation of how well the disruption indicators work. Whenever observational or nonexperimental data are available in scientometrics, it should be considered that causality cannot be proven statistically. Instead, one has to rely on

theoretical reasoning for identifying relevant confounders (citation impact in our case). If one requires causal interpretation of the results, it is necessary to explain and outline plausibly that all potential confounding factors are accounted for. This necessity applies independently of the technique used and neither regular regression models nor matching are able to "prove" causality through statistics. When the number of potential confounders is large in a scientometric study, it is possible to generate compound indices by using methods of data reduction. However, if central confounding factors are not available in the data or cannot be included in the matching process (such as citation impact in our case), one should refrain from causal interpretations. Thus, with respect to our data set, we are not able to interpret the computed statistics (ATEs) as unbiased causal effects but rather as associations.

6. TAKE-HOME MESSAGES

In this section, we summarize the most crucial aspects of using matching techniques in scientometric studies.

- Start by building your theoretical framework and formulate testable hypotheses. Name all potential confounders and describe how they are measured. When not all relevant confounders are measured, refrain from a causal interpretation of the results.
- Compute descriptive statistics for all dependent and independent variables you are going to use. The statistics help to choose the correct models and operationalization. For example, when mostly categorical covariates with few categories are used, exact matching might be a good solution. However, when the number of categories is large or continuous outcomes are utilized, tests of different algorithms to group these variables for CEM can be beneficial. Try to find a good balance between a reduction of bias and the number of cases left for analyses.
- Compute results for various matching techniques. This can be the most crucial aspect of the analyses because most techniques come with a large number of options. As these options also depend on the software package used to compute the results, inspecting the documentation is highly relevant. Either programmers themselves give recommendations for how to use certain options or you should test how strongly outcomes diverge when different options are utilized.
- Inspect balancing for each analysis. Report the results that minimize imbalance or report all results for comparison. Make sure to report balancing either using tables or graphics. If the balancing displays larger deviations between treatment and control group even after matching, the results might not be trustworthy and biases could be present. If the deviations are small, this does not mean that the results are unbiased (omitted variables could still have a confounding effect). It might show that balance is achieved between treatment and control group and there is no confounding left with respect to all control variables used in the models. This can be an iterative procedure: Insufficient balancing requires tweaking the matching model until sufficient balance is reached.
- Compute both regular (analytical) and bootstrapped standard errors for all coefficients of interest (e.g., ATE). The rationale behind this is to rely on two quite different assumptions. Regular standard errors are parametric and depend on the assumption of normality, which is rather strong. Bootstrapped standard errors require fewer assumptions but more computational effort. When both standard errors come to similar conclusions, this points to the stability of the findings. If deviations between the standard errors are large, it should be checked whether there are underlying problems with the models or

variables used. This might concern the skewness of continuous variables that deviates from the normal distribution. If no such obvious problems can be detected, report both types of standard errors and acknowledge that the results are potentially not very stable.

- When reporting the empirical findings, be transparent and describe the details of your results (software used, matching algorithms, imbalance, type of standard errors, etc.). Provide the source code (and raw data, if allowed) to aid replication studies.
- Use regression models as an additional robustness check. Regression techniques are highly popular for good reasons and reach beyond what matching can offer at the moment (for example, the consideration of various outcome variables and link families).

ACKNOWLEDGMENTS

We would like to thank two anonymous reviewers for their detailed and helpful comments and Ben Jann for information on matching algorithms and his strategy for implementing them in Stata.

AUTHOR CONTRIBUTIONS

Felix Bittmann: Methodology, Software, Formal analysis, Writing—original draft, Writing—review & editing, Visualization. Alexander Tekles: Software, Data curation, Writing—review & editing. Lutz Bornmann: Conceptualization, Supervision, Project administration, Writing—review & editing.

COMPETING INTERESTS

The authors have no competing interests.

FUNDING INFORMATION

The authors received no financial support for the research, authorship, and/or publication of this article.

DATA AVAILABILITY

The data cannot be made openly available, as this is not allowed by the data provider Clarivate Analytics.

REFERENCES

Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, *105*(490), 493–505. https://doi.org/10.1198/jasa.2009.ap08746

Abadie, A., & Imbens, G. W. (2016). Matching on the estimated propensity score. *Econometrica*, *84*(2), 781–807. https://doi.org/10.3982/ECTA11293

Ahlgren, P., Colliander, C., & Sjögårde, P. (2018). Exploring the relation between referencing practices and citation impact: A large-scale study based on Web of Science data. *Journal of the Association for Information Science and Technology*, *69*(5), 728–743. https://doi.org/10.1002/asi.23986

Amusa, L., Zewotir, T., & North, D. (2019). Examination of entropy balancing technique for estimating some standard measures of treatment effects: A simulation study. *Electronic Journal of Applied Statistical Analysis*, *12*(2), 491–507.

Austin, P. C., & Cafri, G. (2020). Variance estimation when using propensity-score matching with replacement with survival or time-to-event outcomes. *Statistics in Medicine*, *39*(11), 1623–1640. https://doi.org/10.1002/sim.8502, PubMed: 32109319

Austin, P. C., & Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine*, *34*(28), 3661–3679. https://doi.org/10.1002/sim.6607, PubMed: 26238958

Beaver, D. B. (2004). Does collaborative research have greater epistemic authority? *Scientometrics*, *60*(3), 399–408. https://doi.org/10.1023/B:SCIE.0000034382.85360.cd

Bittmann, F. (2019). *Stata: A really short introduction*. Munich: De Gruyter Oldenbourg. https://doi.org/10.1515/9783110617160

Bittmann, F. (2021). *Bootstrapping. An integrated approach with Python and Stata*. Munich: De Gruyter Oldenbourg. https://doi.org/10.1515/9783110693348

Bornmann, L., Devarakonda, S., Tekles, A., & Chacko, G. (2019). Are disruption index indicators convergently valid? The comparison of several indicator variants with assessments by peers. *Quantitative Science Studies*, *1*(3), 1242–1259. https://doi.org/10.1162/qss_a_00068

Bornmann, L., Devarakonda, S., Tekles, A., & Chacko, G. (2020). Disruptive papers published in *Scientometrics*: Meaningful results by using an improved variant of the disruption index originally proposed by Wu, Wang and Evans (2019). *Scientometrics*, *123*(2), 1149–1155. https://doi.org/10.1007/s11192-020-03406-8

Bornmann, L., & Tekles, A. (2019). Disruption index depends on length of citation window. *El profesional de la información*, *28*(2), e280207. https://doi.org/10.3145/epi.2019.mar.07

Bornmann, L., & Tekles, A. (2021). Convergent validity of several indicators measuring disruptiveness with milestone assignments to physics papers by experts. *Journal of Informetrics*.

Bu, Y., Waltman, L., & Huang, Y. (2021). A multi-dimensional framework for characterizing the citation impact of scientific publications. *Quantitative Science Studies*, *2*(1), 155–183. https://doi.org/10.1162/qss_a_00109

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, *22*(1), 31–72. https://doi.org/10.1111/j.1467-6419.2007.00527.x

Canty, A., & Ripley, B. (2021). *boot: Bootstrap R (S-Plus) functions*. R package version 1.3-28.

D'Agostino Jr., R. B. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine*, *17*(19), 2265–2281. https://doi.org/10.1002/(SICI)1097-0258(19981015)17:19<2265::AID-SIM918>3.0.CO;2-B, PubMed: 9802183

DiPrete, T. A., & Gangl, M. (2004). Assessing bias in the estimation of causal effects: Rosenbaum bounds on matching estimators and instrumental variables estimation with imperfect instruments. *Sociological Methodology*, *34*(1), 271–310. https://doi.org/10.1111/j.0081-1750.2004.00154.x

Doane, D. P. (1976). Aesthetic frequency classifications. *The American Statistician*, *30*(4), 181–183. https://doi.org/10.1080/00031305.1976.10479172

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. New York: CRC Press. https://doi.org/10.1201/9780429246593

Farys, R., & Wolbring, T. (2017). Matched control groups for modeling events in citation data: An illustration of Nobel prize effects in citation networks. *Journal of the Association for Information Science and Technology*, *68*(9), 2201–2210. https://doi.org/10.1002/asi.23802

Fok, D., & Franses, P. H. (2007). Modeling the diffusion of scientific publications. *Journal of Econometrics*, *139*(2), 376–390. https://doi.org/10.1016/j.jeconom.2006.10.021

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., ... Barabási, A.-L. (2018). Science of science. *Science*, *359*(6379), eaao0185. https://doi.org/10.1126/science.aao0185, PubMed: 29496846

Frölich, M. (2007). On the inefficiency of propensity score matching. *Advances in Statistical Analysis*, *91*(3), 279–290. https://doi.org/10.1007/s10182-007-0035-0

Funk, M. D., Westreich, D., Wiesen, C., Stürmer, T., Brookhart, M. A., & Davidian, M. (2011). Doubly robust estimation of causal effects. *American Journal of Epidemiology*, *173*(7), 761–767. https://doi.org/10.1093/aje/kwq439, PubMed: 21385832

Funk, R. J., & Owen-Smith, J. (2017). A dynamic network measure of technological change. *Management Science*, *63*(3), 791–817. https://doi.org/10.1287/mnsc.2015.2366

Gingras, Y., & Khelfaoui, M. (2018). Assessing the effect of the United States' "citation advantage" on other countries' scientific impact as measured in the Web of Science (WoS) database. *Scientometrics*, *114*(2), 517–532. https://doi.org/10.1007/s11192-017-2593-6

Ginther, D. K., & Heggeness, M. L. (2020). Administrative discretion in scientific funding: Evidence from a prestigious postdoctoral training program. *Research Policy*, *49*(4). https://doi.org/10.1016/j.respol.2020.103953, PubMed: 32675837

Guarcello, M. A., Levine, R. A., Beemer, J., Frazee, J. P., Laumakis, M. A., & Schellenberg, S. A. (2017). Balancing student success: Assessing supplemental instruction through coarsened exact matching. *Technology, Knowledge and Learning*, *22*(3), 335–352. https://doi.org/10.1007/s10758-017-9317-0

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, *20*(1), 25–46. https://doi.org/10.1093/pan/mpr025

Hainmueller, J. (2014). *ebal: Entropy reweighting to create balanced samples. R package version 0.1-6*. https://CRAN.R-project.org/package=ebal

Halpern, E. F. (2014). Behind the numbers: Inverse probability weighting. *Radiology*, *271*(3), 625–628. https://doi.org/10.1148/radiol.14140035, PubMed: 24848956

Heinrich, C., Maffioli, A., & Vazquez, G. (2010). *A primer for applying propensity-score matching. Inter-American Development Bank*. Retrieved 5 February 2021, from https://publications.iadb.org/publications/english/document/A-Primer-for-Applying-Propensity-Score-Matching.pdf

Hill, J. (2008). Discussion of research using propensity-score matching: Comments on 'A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003' by Peter Austin, Statistics in Medicine. *Statistics in Medicine*, *27*(12), 2055–2061. https://doi.org/10.1002/sim.3245, PubMed: 18446836

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, *42*(8), 1–28. https://doi.org/10.18637/jss.v042.i08

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, *47*(260), 663–685. https://doi.org/10.1080/01621459.1952.10483446

Iacus, S. M., King, G., & Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political Analysis*, *20*(1), 1–24. https://doi.org/10.1093/pan/mpr013

Imbens, G. W., & Rubin, D. B. (2015): *Causal inference for statistics, social, and biomedical sciences: An introduction*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139025751

Jann, B. (2017a). *KMATCH: Stata module for multivariate-distance and propensity score matching, including entropy balancing, inverse probability weighting, (coarsened) exact matching and regression adjustment*. Retrieved 5 February 2021, from https://ideas.repec.org/c/boc/bocode/s458346.html

Jann, B. (2017b). *Why propensity scores should be used for matching. German Stata Users Group Meeting, Berlin*. https://doi.org/10.7892/boris.101593

Jann, B. (2019). Influence functions for linear regression (with an application to regression adjustment). Retrieved 5 February 2021, from https://doi.org/10.7892/boris.130362

King, G., & Nielsen, R. (2019). Why propensity scores should not be used for matching. *Political Analysis*, *27*(4), 435–454. https://doi.org/10.1017/pan.2019.11

Milojević, S. (2013). Accuracy of simple, initials-based methods for author name disambiguation. *Journal of Informetrics*, *7*(4), 767–773. https://doi.org/10.1016/j.joi.2013.06.006

Morgan, S. L., & Winship, C. (2015). *Counterfactuals and causal inference: Methods and principles for social research* (2nd Edn). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781107587991

Mutz, R., & Daniel, H.-D. (2012). Skewed citation distributions and bias factors: Solutions to two core problems with the journal impact factor. *Journal of Informetrics*, *6*(2), 169–176. https://doi.org/10.1016/j.joi.2011.12.006

Mutz, R., Wolbring, T., & Daniel, H.-D. (2017). The effect of the "very important paper" (VIP) designation in *Angewandte Chemie International Edition* on citation impact: A propensity score matching analysis. *Journal of the Association for Information Science and Technology*, *68*(9), 2139–2153. https://doi.org/10.1002/asi.23701

Olmos, A., & Govindasamy, P. (2015). Propensity scores: A practical introduction using R. *Journal of MultiDisciplinary Evaluation*, *11*(25), 68–88. https://journals.sfu.ca/jmde/index.php/jmde_1/article/download/431/414

Onodera, N., & Yoshikane, F. (2015). Factors affecting citation rates of research articles. *Journal of the Association for Information Science and Technology*, *66*(4), 739–764. https://doi.org/10.1002/asi.23209

Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys*, *3*, 96–146. https://doi.org/10.1214/09-SS057

Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal inference in statistics: A primer*. Hoboken, NY: Wiley.

Pearl, J., & Mackenzie, D. (2019). *Book of why: The new science of cause and effect*. New York: Basic Books.

Peters, H. P., & van Raan, A. F. (1994). On determinants of citation scores: A case study in chemical engineering. *Journal of the American Society for Information Science*, *45*(1), 39–49. https://doi.org/10.1002/(SICI)1097-4571(199401)45:1<39::AID-ASI5>3.0.CO;2-Q

PRE Milestones. (n.d.). *Physical Review E*. Retrieved 11 August 2020, from https://journals.aps.org/pre/collections/pre-milestones

Randolph, J. J., Falbe, K., Manuel, A. K., & Balloun, J. L. (2014). A step-by-step guide to propensity score matching in R. *Practical Assessment, Research & Evaluation*, *19*(18).

Rosenbaum, P. R. (1999). Choice as an alternative to control in observational studies. *Statistical Science*, *14*(3), 259–278. https://doi.org/10.1214/ss/1009212410

Rosenbaum, P. R. (2002). *Observational studies*. New York: Springer. https://doi.org/10.1007/978-1-4757-3692-2

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, *39*(1), 33–38. https://doi.org/10.1080/00031305.1985.10479383

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. https://doi.org/10.1037/h0037350

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine*, *26*(1), 20–36. https://doi.org/10.1002/sim.2739, PubMed: 17072897

Schurer, S., Alspach, M., MacRae, J., & Martin, G. (2016). The medical care costs of mood disorders: A coarsened exact matching approach. *Economic Record*, *92*(296), 81–93. https://doi.org/10.1111/1475-4932.12218

Seglen, P. O. (1992). The skewness of science. *Journal of the American Society for Information Science*, *43*(9), 628–638. https://doi.org/10.1002/(SICI)1097-4571(199210)43:9<628::AID-ASI5>3.0.CO;2-0

Stevens, G. A., King, G., & Shibuya, K. (2010). Deaths from heart failure: Using coarsened exact matching to correct cause-of-death statistics. *Population Health Metrics*, *8*(1), 1–9. https://doi.org/10.1186/1478-7954-8-6, PubMed: 20388206

Thoemmes, F. (2012). *Propensity score matching in SPSS*. Retrieved 5 February 2021, from https://arxiv.org/pdf/1201.6385.pdf

Tregenza, T. (2002). Gender bias in the refereeing process? *Trends in Ecology & Evolution*, *17*(8), 349–350. https://doi.org/10.1016/S0169-5347(02)02545-4

Valderas, J. M., Bentley, R. A., Buckley, R., Wray, K. B., Wuchty, S., … Uzzi, B. (2007). Why do team-authored papers get cited more? *Science*, *317*(5844), 1496–1498. https://doi.org/10.1126/science.317.5844.1496b, PubMed: 17872425

van Wesel, M., Wyatt, S., & ten Haaf, J. (2014). What a difference a colon makes: How superficial factors influence subsequent citation. *Scientometrics*, *98*(3), 1601–1615. https://doi.org/10.1007/s11192-013-1154-x

Wei, C., Zhao, Z., Shi, D., & Li, J. (2020). *Nobel-prize-winning papers are significantly more highly-cited but not more disruptive than non-prize-winning counterparts*. Retrieved 5 February 2021, from https://www.ideals.illinois.edu/bitstream/handle/2142/106575/Contribution_477_final.pdf

Wu, L., Wang, D., & Evans, J. A. (2019). Large teams develop and small teams disrupt science and technology. *Nature*, *566*(7744), 378–382. https://doi.org/10.1038/s41586-019-0941-9, PubMed: 30760923

Yu, T., & Yu, G. (2014). Features of scientific papers and the relationships with their citation impact. *Malaysian Journal of Library & Information Science*, *19*(1), 37–50.

Zhao, Q., & Percival, D. (2016). Entropy balancing is doubly robust. *Journal of Causal Inference*, *5*(1), 20160010. https://doi.org/10.1515/jci-2016-0010