

Estimating the Causal Effect of Early ArXiving on Paper Acceptance

Yanai Elazar^{1,2*} Jiayao Zhang^{3*} David Wadden^{1*} Bo Zhang⁴ Noah A. Smith^{1,2}

¹Allen Institute for Artificial Intelligence

²Paul G. Allen School of Computer Science & Engineering, University of Washington

³University of Pennsylvania

⁴Fred Hutchinson Cancer Center

yanaiela@gmail.com jiayaozhang@acm.org davidw@allenai.org

Editors: Francesco Locatello and Vanessa Didelez

Abstract

What is the effect of releasing a preprint of a paper before it is submitted for peer review? No randomized controlled trial has been conducted, so we turn to observational data to answer this question. We use data from the ICLR conference (2018–2022) and apply methods from causal inference to estimate the effect of arXiving a paper before the reviewing period (*early arXiving*) on its acceptance to the conference. Adjusting for confounders such as topic, authors, and quality, we may estimate the causal effect. However, since quality is a challenging construct to estimate, we use the *negative outcome control* method, using *paper citation count* as a control variable to debias the quality confounding effect. Our results suggest that early arXiving may have a small effect on a paper’s chances of acceptance. However, this effect (when existing) does not differ significantly across different groups of authors, as grouped by author citation count and institute rank. This suggests that early arXiving does not provide an advantage to any particular group.¹

Keywords: Causal inference, negative outcome control, peer-review

1. Introduction

In double-blind peer review, the identities of paper authors and reviewers remain hidden. However, due to the rapidly increasing popularity of the online preprint server arXiv² many papers are now available as preprints before they have been submitted to conferences for peer review.

While this process does not directly break double-blind review, the availability of arXiv preprints makes it possible for reviewers to determine the identity of the authors associated with an arXived paper submission. Reviewers, in turn, may be biased based on characteristics of the paper authors such as their citation counts, or their institutions’ rankings. Prior studies comparing single- and double-blind review provide mixed evidence, with some pointing to bias in favor of highly-ranked authors or universities (Tomkins et al., 2017) and others failing to find such a bias (Madden and DeWitt, 2006). Crucially, no prior work has studied the effect of arXiving (publishing on arXiv) before a submission deadline—which may or not break double-blindness in practice—on paper acceptance.

Motivated by these observations, we aim to answer the following questions: **(RQ1)** Does early arXiving result in different effects on authors based on their citation counts and institutions? **(RQ2)** What effect (if any) does early arXiving have on paper acceptance?

* Equal contribution; random ordering. Work done while JZ interned at AI2.

1. Code and data can be found at: <https://github.com/allenai/anonymity-period>.

2. arxiv.org

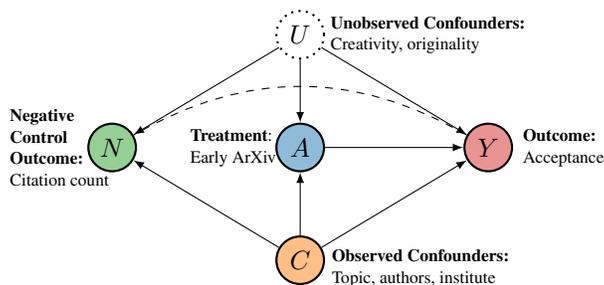


Figure 1: **Causal graph of our problem.** A and Y are binary treatment and effect variables, respectively: whether a paper was arXived before the review deadline, and whether the paper was accepted. As we cannot measure the unobserved confounders (e.g., quality), we estimate the effect of arXiving using a negative control outcome variable (N). Solid edges represent a directed causal effect, while dashed edges represent an association.

To address these questions, we take advantage of a dataset encompassing five years of paper submissions to the ICLR conference. ICLR is unique because it is the only conference which (to our knowledge) releases the acceptance outcomes (accept or reject) of all submitted papers; the availability of these acceptance decisions enables the causal analysis performed in this work.³ We operationalize the problem as a causal graph (Figure 1) where paper acceptance (the *outcome* variable) is determined by three key factors: (1) whether or not the paper was arXived prior to the review deadline (the *treatment* variable), (2) *observed confounders* like paper topic, author citations, and host institution rankings, and (3) *unobserved confounders* like paper quality and potential for impact.

Unobserved confounders present a challenge, since they affect acceptance decisions, but we cannot directly adjust for them since they cannot be straightforwardly measured. To address this challenge, we employ a well-established causal inference technique known as the *negative control outcome* (NCO).⁴ This approach requires identifying an NCO variable which is affected by the same confounders as paper acceptance (our outcome variable), but is *not* affected by early arXiving (our treatment variable). We propose to use a paper’s citation count in the years following its initial release as an NCO. Specifically, we binarize citation count to categorize papers as “highly-cited” and “less-cited”, eliminating minor variations in counts. We conduct multiple analyses to ensure our findings are robust to the number of years over which we count citations, and the binarization threshold used.

Overall, we find that **(RQ1)** under standard assumptions, we observe no statistically significant evidence that the effect of early arXiving on acceptance is different for authors with different citation counts or institution ranks, and **(RQ2)** with additional assumptions, the *effect size* of early arXiving on acceptance is either small (less than 4% in seven out of nine settings), or not statistically significant (in four out of nine settings).

3. iclr.cc

4. Causal inference literature also refers to *negative outcome control* (NOC) as the methodology, and NCO as the outcome variable. In this paper we use NCO for both meanings interchangeably for improved readability.

2. Problem Formulation

In this section, we formalize the paper acceptance process using the framework of causal inference. We aim to characterize the effect of the **treatment** A —a binary variable indicating whether the authors submitted their work to arXiv before the **reviewing deadline**⁵—on the binary **outcome** Y —whether the paper was accepted. We refer to A as “early arXiving” for convenience.

There are a number of factors other than A that may affect whether a paper is accepted. These factors, usually called *confounders* or *covariates*, can be divided into two groups. **Observed confounders** C are attributes that we can operationalize, explicitly measure, and include as variables in our analysis. These include features like paper topic and writing style,⁶ as well as attributes of the paper authors and their institutions. We list the 18 confounders we consider in Table 1. **Unobserved confounders** U are attributes that are challenging or impossible to measure, but which nonetheless may impact the outcome Y as well as the treatment A . Possible unobserved confounders in this work include characteristics like creativity, novelty, and potential impact; we refer to these collectively as paper *quality*. Although we cannot measure these characteristics, we present an approach to debias their effect in Section 4. Finally, our analysis also makes use of paper *citation count*. This variable is used to construct an NCO N , which we discuss in depth in Section 4. N can be associated with the acceptance outcome Y , both because a paper’s acceptance may increase its citation count, and also due to shared confounders such as quality. As such, these variables may be correlated, but this does not affect our estimation discussed in the next section.

3. Causal Inference: Background

We briefly review the potential-outcomes framework, also known as the Rubin causal model (Rubin, 1974, 2005) as it applies to our setting. A comprehensive overview is given by Imbens and Rubin (2015); for discussions on causal graphs, see e.g., Pearl (1995); for a review on causal inference methods for NLP see e.g., Feder et al. (2022); for a discussion on using causal inference on textual data, see e.g., Zhang and Zhang (2022).

The outcome we study in this work is a binary variable indicating whether a submitted paper is accepted at a conference (Y). For a given paper there exist two *potential outcomes*: $Y_{A=1}$ is the acceptance outcome that would occur *if the paper were arXived before the review deadline* — i.e., *if A were 1*. Similarly, $Y_{A=0}$ is the outcome that would occur *if A were 0*. In practice, we observe exactly one of these two outcomes for a given paper: either the authors arXived it early or they did not. The goal of causal inference is to reason about *counterfactuals*: how would a paper’s chances of acceptance have changed if the authors had made the other choice?

The most straightforward way to reason about this counterfactual is with the *average treatment effect* (ATE), defined as:

$$\text{ATE} = \mathbb{E}[Y_{A=1} - Y_{A=0}],$$

which is the treatment effect averaged over all individuals from a population. In our study, we consider a popular alternative, the *average treatment effect on treated* (ATET):

$$\text{ATET} = \mathbb{E}[Y_{A=1} - Y_{A=0} \mid A = 1], \tag{1}$$

5. We deliberately choose the reviewing deadline and not the submission deadline to account for papers posted on arXiv between these dates, effectively de-anonymizing the authors.

6. Text features were shown to be important predictors for the problem of citation prediction (Yogatama et al., 2011).

Variable name	Description	Data type
year	Year in which the paper first became available online.	Date
n_fig	Number of figures appearing in the paper.	Integer
n_ref	Number of references (citations) in the paper.	Integer
n_sec	Number of sections in the paper.	Integer
log_text_length	Logarithm of the number of tokens.	Float
text_ppl	Text perplexity, $2^{-\frac{1}{n} \sum_{i=1}^n \log_2 LM(w_i w_{1:i-1})}$.	Float
topic_cluster	Topic cluster (20 total) assigned to a paper.	Categorical
n_author	Total number of paper authors.	Integer
n_author_female	Number of female paper authors.	Integer
first_author_female	Indicator whether the first author is female.	Binary
any_author_female	Indicator whether there are any female authors.	Binary
no_US_author	Indicator whether there are no US-based authors.	Binary
log_inst_rank_min	Log of lowest rank of institution affiliated with any paper author.	Float
log_inst_rank_avg	Log of average rank of author institutions.	Float
log_inst_rank_max	Log of highest rank of institution affiliated with any paper author.	Float
log_author_cite_min	Citation count of least-cited author in log scale.	Float
log_author_cite_avg	Average author citation count in log scale.	Float
log_author_cite_max	Citation count of most-cited author in log scale.	Float

Table 1: **Observed confounders C included in our analysis.** “Institute rank” is computed by counting the total number of accepted papers at ICLR from a given institution, in the two years prior to each submitted paper. “Topic clusters” include common AI topics such as “transfer learning” and “language models”. Gender information is provided by the authors at the creation of their OpenReview profiles. See Table A.1 for the full list of topics, along with additional details and statistics on all confounders.

which in our case helps to answer *whether early arXiving has created an advantage for the group of submissions that, in fact, arXiv early.*⁷

3.1. Assumptions

Our modeling assumptions are described in the causal graph (Figure 1) and we make several standard identification assumptions: *ignorability*, *positivity*, and *consistency*. In addition, we assume *negative control condition* when using the NCO framework. In what follows we formally define and explain these assumptions.

Assumption 3.1 (Identification Assumptions)

1. *Ignorability*: $\{Y_0, Y_1\} \perp\!\!\!\perp A \mid (C, U)$.
2. *Positivity*: $0 < \mathbb{P}(A = 1 \mid C = c, U = u) < 1$.
3. *Consistency*: $Y_a = Y^{obs}$ if $A^{obs} = a$.
4. *Negative control*: $N_a = N$ for $a = 0, 1$.

Intuitively, *ignorability* assumes the observed and unobserved covariates (C, U) are sufficient to deconfound the effect of the treatment (there are no other unobserved confounders); *positivity* means

⁷ In what follows, we simplify notation from $Y_{A=a}$ to Y_a for brevity when there is no potential ambiguity.

each unit (“submission”) has a non-zero chance of being early arXived and a non-zero chance of not being early arXived; *consistency* establishes the relationship between potential-outcome Y_a and observed outcome Y^{obs} under the observed treatment level A^{obs} : the potential-outcome under $A = a$ agrees with the observed outcome when $A = a$. Finally, we make use of another assumption for dealing with unobserved covariates such as quality; *negative control condition* dictates that the NCO is not causally affected by the treatment.

4. Estimating the Causal Effect

Causal Estimand We write $Y_{A=a}$ as the *potential-outcome* of Y under the treatment $A = a$ for $a = 0, 1$. Our goal is to estimate the average treatment effect on the treated,

$$\text{ATET} = \mathbb{E}[Y_{A=1} - Y_{A=0} \mid A = 1].$$

ATET assesses whether early arXiving has created an advantage for the submissions which were arXived early. The ATET ranges between -100% to 100% , where larger absolute values indicate a stronger effect of the treatment (either positive or negative), whereas values closer to 0 indicate weak to no effect. Our modeling assumptions are detailed in the causal graph (Figure 1). We make several standard identification assumptions (including one for the NCO), discussed in detail in Section 3.1.

4.1. Effect Estimation

To estimate the ATET using NCO, we make use of a common technique known as *Difference-in-Difference* (Lechner, 2011; Sofer et al., 2016). Additionally, we perform statistical matching (Rosenbaum and Rubin, 1983; Rosenbaum, 1989b; Rubin, 2005) to ensure that our treated and control (i.e., early arXiv vs. not) groups are comparable. All of our subsequent analyses are conducted on the matched sample.

Negative Control Outcome and Difference-in-Difference NCO has been studied extensively and used in various domains to debias the effect estimate under the presence of unknown confounding (Rosenbaum, 1989a; Weiss, 2002; Lipsitch et al., 2010; Shi et al., 2022), where NCO variables are usually motivated from the nature of biological mechanisms (Lu et al., 2008). Difference-in-Difference (DiD) is a popular way of estimating ATET in longitudinal studies when the outcome is measured in two time periods $t = 0, 1$; its history dates back to the nineteenth century (Snow, 1854; Card and Krueger, 1994; Bertrand et al., 2004; Lechner, 2011). In this setup four potential-outcomes $Y_a(t)$ for $a, t = 0, 1$ are defined and the ATET can be rewritten as:

$$\text{ATET} = \mathbb{E}[Y_1(1) - Y_0(1)] - \mathbb{E}[Y_1(0) - Y_0(0)].$$

A linear outcome model is usually assumed in the DiD literature, which implies the following condition that relates the measurements over time:

Assumption 4.1 (Additive Equi-confounding)

$$\mathbb{E}[Y_a(1) - Y_a(0) \mid U, A = a, C] = \mathbb{E}[Y_a(1) - Y_a(0) \mid A = a, C],$$

for $A = 0, 1$.

Under this assumption, ATET can be estimated through taking the difference (over treatment levels) of the difference (over time periods), hence its name “difference-in-difference”. In the case of binary outcome variables, one may assume $Y_a(t)$ follows a logistic model. The ATET can then be estimated by computing the conditional means (i.e., using the fitted model to predict the probability that $Y_a(t)$ assumes a certain value), and confidence intervals can be estimated through bootstrapping.

Sofer et al. (2016) establish the equivalence between NCO and DiD-based estimators by inspecting their identification assumptions and note that the role of an NCO variable N is equivalent to $Y(0)$ (the outcome at the first time period $t = 0$), thus the ATET can be rewritten, in the case for an NCO N , as

$$\text{ATET} = \mathbb{E}[Y_1 - N_1] - \mathbb{E}[Y_0 - N_0].$$

Given the additive equi-confounding assumption that Y and N are affected by U in a similar way, a debiased ATET can be estimated by taking the difference (across the treated and control groups) of the difference between Y and N . Combining statistical matching and DiD renders our analysis more robust to model misspecifications (Rubin, 1979).

Matching We use statistical matching (Rosenbaum and Rubin, 1983; Rosenbaum, 1989b; Rubin, 2005) to ensure our observed covariates (listed in Table 1) are *balanced* in the treated and matched control groups, thus making the two groups comparable. We apply matching to obtain a matched pair in the control group for each submission in the treated group. This ensures each matched pair is comparable in terms of covariates and only differs by the treatment level. We follow Chen et al. (2022) in using the tripartite matching algorithm (Zhang et al., 2021), to ensure (1) numerical variables are matched (with a penalty on their L_2 difference); (2) categorical variables `n_author` and `year` are nearly exactly matched; (3) the distribution of `topic_clusters` in the matched group is similar to that in the treated group. After matching, all 1,486 treated units are matched, and the discrepancy between covariates across treated and matched control groups are significantly reduced, as shown in Table A.1 in the Appendix.

4.2. Choosing a Negative Control Outcome Variable

Paper quality is a crucial unobserved confounder for paper acceptance decisions. We use the NCO framework (Rosenbaum, 1989a; Lipsitch et al., 2010) to adjust for the confounding effect of the unobserved confounders U . By identifying a variable N that shares the same observed and unobserved confounders as our outcome Y , while being not *causally affected* by the treatment A ⁸ (also entailed by Figure 1), we can then attempt to correct the bias due to unobserved confounders.

In this study, we define an NCO, denoted $N_q^{(n)}$, based on a paper’s citation count in the n years following its *initial* release, denoted as $\text{CC}^{(n)}$.⁹ **Importantly**, we retrieve the date of each paper’s *first* public appearance—whether on arXiv, in conference proceedings, or some other source (see Appendix A.2 for details).¹⁰ Measuring citations in a fixed time window since each paper first became available online—rather than by calendar year or starting from the date of the conference in which the paper was published—eliminates a source of confounding where early-arXived papers would have higher values for $\text{CC}^{(n)}$ simply because they were available earlier.

8. $N \perp\!\!\!\perp A \mid C, U$

9. Note that citation count is *not* to be confused with a measure of paper quality.

10. For instance, if a paper first appeared publicly as an arXiv preprint on May 15, 2018, then $\text{CC}^{(2)}$ for this paper counts the number of citations as of May 15, 2020.

For a given paper, $N_q^{(n)} = 1$ if the paper’s citation count n years after its release is above the q th quantile of citation counts for all papers in the relevant sample, $CC_q^{(n)}$. Intuitively, $N_q^{(n)}$ captures whether a given paper was “highly-cited” or “less-cited” after n years. In Section 5, we conduct analyses using values of $\{1, 2, 3\}$ for n , and $\{50\%, 75\%, 90\%\}$ for q .

Validity of Citation Count as a Negative Control Outcome By choosing $N_q^{(n)}$ as our NCO, we are assuming that whether or not a paper happened to be arXived early does not have an appreciable effect on the paper’s citation count in a fixed time window following its appearance. We address two possible concerns with this assumption.

First, prior work by Feldman et al. (2018) has observed an association between early arXiving and increased citations in the calendar year of the paper’s release. These findings do not contradict our NCO assumption, for two reasons: (1) Feldman et al. (2018) recognize that their findings are not causal, since they do not account for the presence of unobserved confounders in their analysis, such as the paper’s quality. (2) The choice to use citations in the calendar year of publication was necessitated by the lack of availability of exact publication dates in the Semantic Scholar search engine¹¹ at the time the work was performed. Fortunately, exact publication dates are now available, eliminating the potential confounding effects of measuring citations by calendar year.

Second, researchers (Feldman et al., 2018; Goldberg, 2017) have voiced concerns that arXiving can be used as a technique for “flag-planting”, whereby authors hope to take credit for an idea (and receive more citations) by posting it on arXiv, while other researchers with the same idea wait for their papers to be reviewed. While this is a plausible concern, we are unaware of any work that has empirically demonstrated a flag-planting effect, and are skeptical that such an effect would be so widespread as to invalidate citation count as an NCO. We welcome an empirical study to quantify the effect of flag-planting in future work. Fortunately, should an even more suitable NCO become available, it can be easily substituted into the causal framework proposed in this work.

5. The Effect of arXiving on Acceptance

We begin by describing the data used for our study (§5.1). Then, we estimate the effect of early arXiving on paper acceptance while controlling for observed confounders, but *not accounting* for unobserved confounders (§5.2). This provides us with a so-called *primary analysis* to assess effect sizes in the matched sample, without relying on any NCO assumptions. We then stratify the analysis by institution ranking and author citation count to address **RQ1** (§5.3). Finally, by accounting for the unobserved confounders using NCO (§4) we seek an answer for **RQ2** (§5.4).

5.1. Dataset

We use the ICLR 2018–2022 database assembled by Zhang et al. (2022), which includes 10,297 papers. We select seven submission-related covariates and eleven author-related covariates, shown in Table 1. For each of 1,486 early arXived papers, we find a matched non-early arXived paper by matching on all observed confounders. Additional details including the summary of covariates before and after matching (Table A.1) are given in Appendix A. When using citation counts, recent publications where $CC^{(n)}$ values are undefined are discarded. For example, when using $CC^{(3)}$, only conferences in or before 2020 ($2023 - 3 = 2020$) are kept.

11. [semanticscholar.org](https://www.semanticscholar.org)

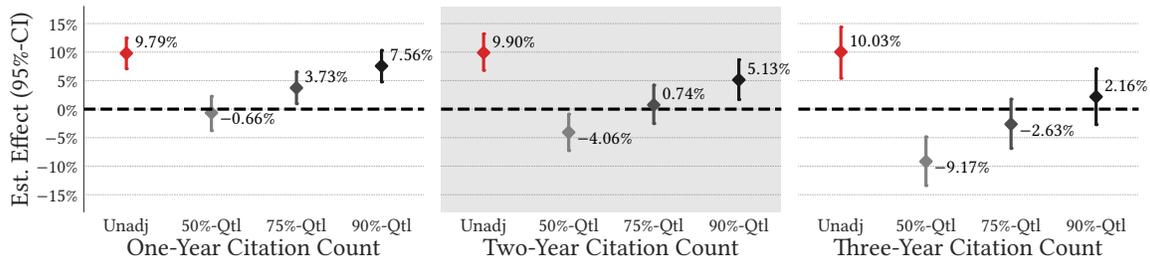


Figure 2: **Effects of early arXiving (with 95% bootstrap confidence interval) estimated on the matched sample.** “Unadj” refers to the estimate without using any NCOs, on the same subset of data that are used by the NCO in the same panel. NCOs are defined to be whether n -year citation is greater than the q th quantile for $n = 1, 2, 3$ and $q \in \{0.5, 0.75, 0.9\}$. Estimated effects without debiasing using NCOs are shown in red. Note that effects estimated using $N_{0.5}^{(1)}$, $N_{0.75}^{(2)}$, $N_{0.9}^{(3)}$, and $N_{0.9}^{(3)}$ are insignificant at the 95% level (confidence intervals contain 0); and effects, where significant, are reduced compared with their non-adjusted counterparts (marked in red). This indicates that NCOs help to explain a large part of unbiased effects.

5.2. Primary Analysis: Controlling for Observed Confounders

We show in Figure 2 the estimated effects of early arXiving on the matched sample. Effect estimates from the primary analysis (i.e., controlling only for observed confounders, while not accounting for unobserved confounders) are marked in red (note that the values are different since the sample changes as different n -year citation counts are used). The primary analysis on the unmatched groups indicates a significant association between early arXiving and paper acceptance, and the effect size is relatively high: 9.79%, 9.90%, and 10.03% for the samples from the different groups. However, in the matched samples from the same groups, early arXived papers are cited almost 2.13 times more (in a three-year window) than their non early-arXived matches. This suggests that a strong confounder such as paper quality likely exists. We thus attempt to use NCO to study whether the statistically-significant effect of early arXiving effect is attributable to the unobserved confounders. In what follows we conduct analyses using our data to answer **RQ1** and **RQ2**.

5.3. Analysis on Author Subgroups

One of the concerns regarding the anonymity period in conferences is that arXiving papers may confer an unfair advantage to authors from prestigious institutions, or who are well-known in the community; see [Eisner et al. \(2017\)](#) for ACL Guidelines. We acknowledge these aspects have little to do with quality, but are often thought to be biasing factors during the review process ([Tomkins et al., 2017](#); [Snodgrass, 2006](#)). Based on our data and assumptions, *do we have evidence that authors from different groups are treated differently if they early arXiv (RQ1)?*

Figure 3 shows estimated effects grouped by institution ranking and author citation count (taking min/max respectively in the case of multiple authors). The lengths of the confidence bands differ, since the set of covariates and sample sizes differ across groups (see Appendix B for details).

In the primary analysis, without any attempts to adjust for unobserved confounders, we do not find evidence that early arXiving is uniquely beneficial to authors belonging to any specific group, as evidenced by the overlapping confidence intervals for all authors groups (‘Unadj’, first row of

ESTIMATING THE CAUSAL EFFECT OF EARLY ARXIVING ON PAPER ACCEPTANCE

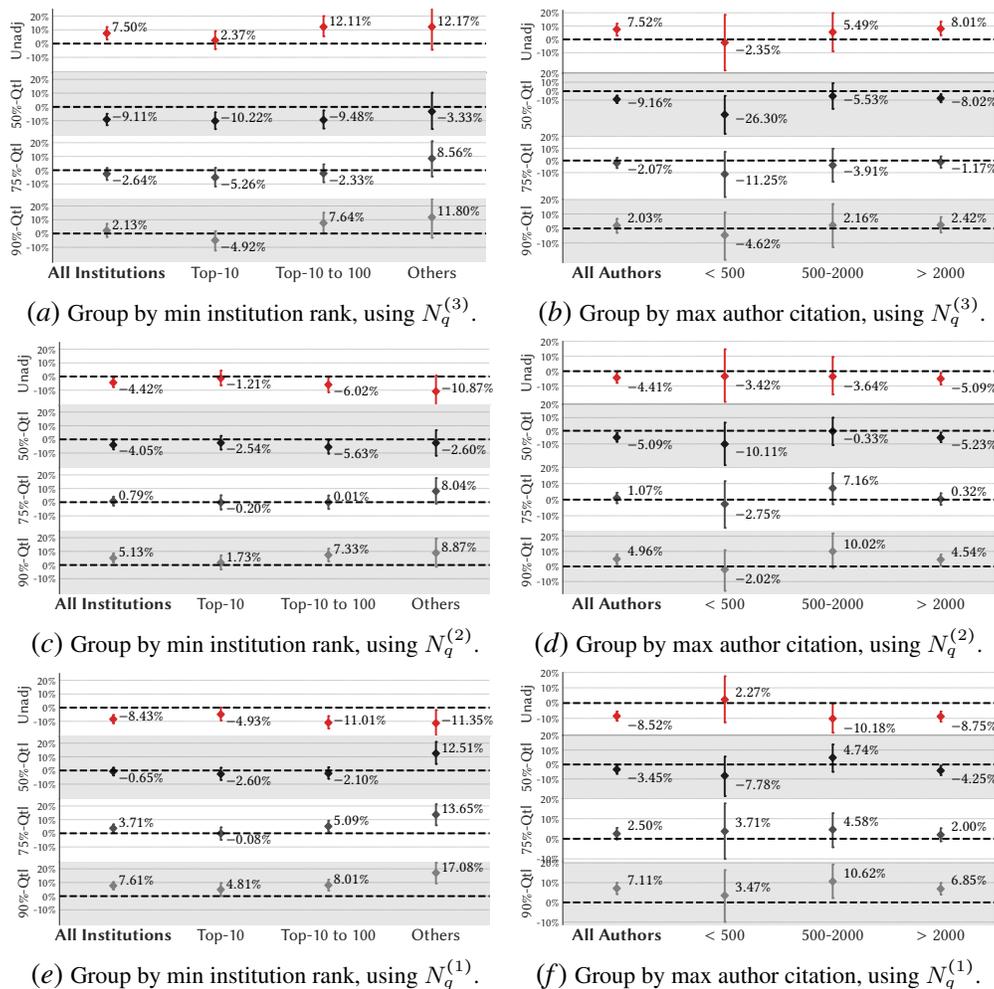


Figure 3: **Estimated effects in author subgroups on the matched sample.** We use $N^{(n)}$, with three values for $n \in \{1, 2, 3\}$ as the NCO and estimate the effect across submissions grouped by the minimum author institution and maximum author citation in the submission (“Unadj” refers to the estimate without using any NCO). Note that for NCOs defined using 75% and 90% quantiles, effects are insignificant (confidence intervals contain 0), and there is no evidence that the effects differ across subgroups (confidence intervals overlap).

each graph). For instance, when using N_q^3 and not adjusting for unobserved confounders, the ATET across all institutions (Figure 3(a)) is 7.5%; by group, the values are 2.37%, 12.11%, and 12.17% for the top-10, top-11 to 100, and all others, respectively. All confidence intervals overlap, meaning we do not observe statistically significant differences between the different groups. The same can be said when we use NCOs to debias the estimate, e.g., for the 90% quantile when stratifying between authors citations when using N_q^3 (Figure 3(b)), the ATET is 2.03%, -4.62%, 2.16%, and 2.42% for all authors, authors with less than 500 citations, between 500-2,000 citations, and more than 2,000 citations, respectively. These analyses indicate that we do not have evidence that arXiving confers a distinct advantage (in terms of paper acceptance) to any particular author subgroup we consider.

5.4. NCO Analysis: Accounting for Unobserved Confounders

Finally, we estimate the ATET and 95%-bootstrap confidence intervals when the NCO $N_q^{(n)}$ is used in an attempt to adjust for unobserved confounders like paper quality. The results are shown in Figure 2 (the non-red results). We observe that (1) debiased effects are reduced compared with the estimates when not adjusting for unobserved confounders (marked as “Unadj” in the left column), indicating that the simple primary analysis is likely to be confounded and using an NCO helps to deconfound. (2) Some estimated effects are significant at the 95%-level, since their 95%-confidence intervals do not contain 0. This indicates that there are remaining effects that cannot be explained by the NCO we chose. (3) Arguably, the strongest choice of NCO is an $N_q^{(n)}$ with a larger n and higher q , which quantifies whether a paper is highly cited over a longer period of time, and thus is less likely to be affected by early arXiving. As such, we observe that using $N_{0.75}^{(3)}$ and $N_{0.9}^{(3)}$ result in effect sizes of -2.63% and 2.16% , respectively, which are not significant at the 95% level, indicating the remaining effects are weak.

6. Conclusion & Discussion

In this work, we investigate whether arXiving a paper before a conference review deadline influences the likelihood that the paper will be accepted at that conference.

We begin with a primary analysis, which assesses the effect of early arXiving on paper acceptance while controlling for observed confounders, but ignoring unobserved confounders such as paper quality. The results of the primary analysis suggest that early arXiving increases a paper’s chances of acceptance by roughly 10% on average, a statistically significant effect.

However, when we stratify our analysis across different author cohorts grouped by institution ranking and citation count, we do *not* observe a statistically significant difference in the effect of early arXiving across subgroups. Thus, we do not find evidence that early arXiving provides an unfair advantage to any particular group of researchers, one of the arguments given for the anonymity period of conferences.

Next, in an attempt to account for the effects of unobserved confounders like paper quality, we leverage the NCO framework, and argue for the use of paper citation count after n years as the NCO variable. Using NCO, under a range of reasonable assumptions, we find that early arXiving has a substantially weaker effect on acceptance relative to our primary analysis—increasing acceptance likelihood by less than 4% in seven out of nine experimental settings (Figure 2). This suggests that papers which were arXived early differ in important but difficult-to-quantify ways from papers which were not, and that our casual inference procedure accounted for some of these differences.

In summation, we find that early arXiving may have a small effect on the likelihood of paper acceptance, under standard assumptions, but that this effect is constant across different researcher subpopulations and does not disadvantage any particular group. This latter finding is noteworthy because one of the most compelling justifications for conference anonymity periods is that they may prevent particular groups of researchers from gaining an unfair advantage by releasing early preprints of their work. Our findings call this justification into question. Therefore, we would advocate for a randomized controlled trial to examine the effects of early arXiving in a more controlled setting; if such a trial recapitulates the results found here, it would suggest that anonymity periods may not be necessary to preserve fairness across different researcher populations.

Epilogue

On January 12, 2024, the day we received the acceptance notification of this paper at CLeaR 2024, the Association for Computational Linguistics decided to remove the anonymity period policy effective immediately from all of its venues.¹² This paper was greatly influenced by the ACL policy and our curiosity about whether the policy’s original goals and motivations were held in practice. We are unlikely to know whether this paper contributed to this policy lift from ACL (we made the paper first available on June 24, 2023). As with many real-world causal questions, assessing the true causal effect is hard.

Ethics Statement

Part of the decision for starting the anonymity period was motivated by the negative biases that arXiving might have on the double-blind review process (Eisner et al., 2017). Our results suggest that the indirect violation of this process through arXiving does not affect different groups (stratified by institute or citations) differently. Together with additional assumptions, we observe that the effect (when existing) is small (in seven out of nine settings it is less than 4%). As such, our results should not be taken as a ground truth, but as some initial evidence that anonymity periods (such as the ACL anonymity period) may not achieve one of its main goals.

Limitations

This work has a number of limitations. First, we analyze decisions from a single machine learning conference where acceptance decisions are readily available, and thus our results may not generalize to other venues or research communities. Second, the validity of the modeling techniques used in this work rests on assumptions that we believe are plausible, and make the best use of the available data, but which are certainly up for debate. We attempted to control for as many observed covariates as we could, and observed the effects estimated using NCOs drops compared with the effects estimated without any NCOs. Although this suggests that NCOs help to deconfound the effects, the NCO assumption that a long-term (e.g., 3-year) citation count is not causally affected by early arXiving is uncheckable, and the conclusions are contingent on this assumption. We welcome future works to estimate such effects with other assumptions. Ideally, one would conduct a randomized controlled trial (or a randomized encouragement trial) in order to obtain an unbiased estimate of early arXiving effects. We note that such an experiment is not trivial to conduct, as it requires a large enough sample group, consent to participate in an experiment that would randomly assign them to a group and require them to either (1) keep their publication anonymous until the decisions, or (2) arXiv their papers before the reviews.

Acknowledgments

We would like to thank the anonymous reviewers, Ashish Sabharwal, and Tim Althoff, as well as the AllenNLP and Semantic Scholar teams for feedback on drafts of this paper.

12. https://aclweb.org/adminwiki/index.php/ACL_Anonymity_Policy.

References

- Susan Athey and Guido W Imbens. Identification and inference in nonlinear difference-in-differences models. *Econometrica*, 74(2):431–497, 2006.
- Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv:2004.05150*, 2020.
- Marianne Bertrand, Esther Duflo, and Sendhil Mullainathan. How Much Should We Trust Differences-In-Differences Estimates?*. *The Quarterly Journal of Economics*, 119(1):249–275, 02 2004. ISSN 0033-5533. doi: 10.1162/003355304772839588. URL <https://doi.org/10.1162/003355304772839588>.
- David Card and Alan B Krueger. Minimum wages and employment: A case study of the fast-food industry in new jersey and pennsylvania. *The American Economic Review*, pages 772–793, 1994.
- Chang Chen, Jiayao Zhang, Dan Roth, Ting Ye, and Bo Zhang. Association between author metadata and acceptance: A feature-rich, matched observational study of a corpus of iclr submissions between 2017-2022. *arXiv preprint arXiv:2211.15849*, 2022.
- Steven A. Cholewiak, Panos Ipeirotis, Victor Silva, and Arun Kannawadi. SCHOLARLY: Simple access to Google Scholar authors and citation using Python, 2021. URL <https://github.com/scholarly-python-package/scholarly>.
- Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. SPECTER: Document-level Representation Learning using Citation-informed Transformers. In *ACL*, 2020.
- Jason Eisner, Jennifer Foster, Iryna Guryvech, Marti Hearst, Heng Ji, Lillian Lee, Christopher Manning, Paola Merlo, Yusuke Miyao, Joakim Nivre, Amanda Stent, and Ming Zhou. ACL policies and guidelines for submission, review and citation, 2017. URL https://www.aclweb.org/adminwiki/index.php/File:ACL_Guidelines_for_Submission,_Review_and_Citation.pdf.
- Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. Causal Inference in Natural Language Processing: Estimation, Prediction, Interpretation and Beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158, 10 2022. ISSN 2307-387X. doi: 10.1162/tacl_a_00511. URL https://doi.org/10.1162/tacl_a_00511.
- Sergey Feldman, Kyle Lo, and Waleed Ammar. Citation count analysis for papers with preprints. *arXiv preprint arXiv:1805.05238*, 2018.
- Yoav Goldberg. An adversarial review of “adversarial generation of natural language”, 2017. URL <https://medium.com/@yoav.goldberg/an-adversarial-review-of-adversarial-generation-of-natural-language-409ac3378bd7>. Accessed on: 2023-10-27.
- Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

- Rodney Michael Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, Miles Crawford, Doug Downey, Jason Dunkelberger, Oren Etzioni, Robert Evans, Sergey Feldman, Joseph Gorney, David W. Graham, F.Q. Hu, Regan Huff, Daniel King, Sebastian Kohlmeier, Bailey Kuehl, Michael Langan, Daniel Lin, Haokun Liu, Kyle Lo, Jaron Lochner, Kelsey MacMillan, Tyler Murray, Christopher Newell, Smita Rao, Shaurya Rohatgi, Paul L Sayre, Zejiang Shen, Amanpreet Singh, Luca Soldaini, Shivashankar Subramanian, A. Tanaka, Alex D Wade, Linda M. Wagner, Lucy Lu Wang, Christopher Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Madeleine van Zuylen, and Daniel S. Weld. The semantic scholar open data platform. *ArXiv*, abs/2301.10140, 2023.
- Michael Lechner. The estimation of causal effects by difference-in-difference methods. *Foundations and Trends® in Econometrics*, 4(3):165–224, 2011. ISSN 1551-3076. doi: 10.1561/08000000014. URL <http://dx.doi.org/10.1561/08000000014>.
- Marc Lipsitch, Eric J Tchetgen Tchetgen, and Ted Cohen. Negative controls: A tool for detecting confounding and bias in observational studies. *Epidemiology*, 2010.
- Y. Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692, 2019.
- Ying-Jie Lu, Jane Gross, Debby Bogaert, Adam Finn, Linda Bagrade, Qibo Zhang, Jay K. Kolls, Amit Srivastava, Anna Lundgren, Sophie Forte, Claudette M. Thompson, Kathleen F. Harney, Porter W. Anderson, Marc Lipsitch, and Richard Malley. Interleukin-17a mediates acquired immunity to pneumococcal colonization. *PLOS Pathogens*, 4(9):1–11, 09 2008. doi: 10.1371/journal.ppat.1000159. URL <https://doi.org/10.1371/journal.ppat.1000159>.
- Samuel Madden and David J. DeWitt. Impact of double-blind reviewing on sigmod publication rates. *SIGMOD Rec.*, 35:29–32, 2006.
- Andrew Ng, Michael Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001. URL https://proceedings.neurips.cc/paper_files/paper/2001/file/801272ee79cfde7fa5960571fee36b9b-Paper.pdf.
- Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- Paul R. Rosenbaum. The role of known effects in observational studies. *Biometrics*, 45(2):557–569, 1989a. ISSN 0006341X, 15410420. URL <http://www.jstor.org/stable/2531497>.
- Paul R. Rosenbaum. Optimal matching for observational studies. *Journal of the American Statistical Association*, 84(408):1024–1032, 1989b.

- Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.
- Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.
- Donald B Rubin. Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, 74(366a):318–328, 1979.
- Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.
- Xu Shi, Wang Miao, and Eric Tchetgen Tchetgen. A selective review of negative control methods in epidemiology, 2022.
- Richard T. Snodgrass. Single- versus double-blind reviewing: an analysis of the literature. *SIGMOD Rec.*, 35:8–21, 2006.
- John Snow. The cholera near golden-square, and at deptford. *Medical Times and Gazette*, 321(22), 1854.
- Tamar Sofer, David B. Richardson, Elena Colicino, Joel Schwartz, and Eric J Tchetgen Tchetgen. On negative outcome control of unobserved confounding as a generalization of difference-in-differences. *Statistical science : a review journal of the Institute of Mathematical Statistics*, 2016.
- Andrew Tomkins, Min Zhang, and William D. Heavlin. Single versus double blind reviewing at wsdm 2017. *ArXiv*, abs/1702.00502, 2017.
- Noel S. Weiss. Can the "specificity" of an association be rehabilitated as a basis for supporting a causal hypothesis? *Epidemiology*, 13(1):6–8, 2002. ISSN 10443983. URL <http://www.jstor.org/stable/3703241>.
- Dani Yogatama, Michael Heilman, Brendan O’Connor, Chris Dyer, Bryan R Routledge, and Noah A Smith. Predicting a scientific community’s response to an article. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 594–604, 2011.
- Bo Zhang and Jiayao Zhang. Some reflections on drawing causal inference using textual data: Parallels between human subjects and organized texts. In *First Conference on Causal Learning and Reasoning*, 2022. URL <https://openreview.net/forum?id=ZJRRwV41CLz>.
- Bo Zhang, Dylan S. Small, Karen B. Lasater, Matthew McHugh, Jeffrey H. Silber, and Paul R. Rosenbaum. Matching one sample according to two criteria in observational studies. *Journal of the American Statistical Association*, pages 1–12, 2021.
- Jiayao Zhang, Hongming Zhang, Zhun Deng, and Dan Roth. Investigating fairness disparities in peer review: A language model enhanced approach. *ArXiv*, abs/2211.06398, 2022.

ESTIMATING THE CAUSAL EFFECT OF EARLY ARXIVING ON PAPER ACCEPTANCE

Covariate	Early arXiv Papers (Treated Group, n = 1486)	Non-Early arXiv Papers (Unmatched Control, n = 7493)	Matched Comparison Group (n = 1486)	SMD (Before Matching)	SMD (After Matching)
year				0.746	0.002
2018	20 (1.3)	710 (9.5)	20 (1.3)		
2019	48 (3.2)	1146 (15.3)	48 (3.2)		
2020	502 (33.8)	1422 (19.0)	502 (33.8)		
2021	503 (33.8)	1876 (25.0)	502 (33.8)		
2022	413 (27.8)	1967 (26.3)	414 (27.9)		
n_fig	14.3 (7.5)	12.6 (7.3)	14.0 (8.2)	-0.221	0.035
n_ref	44.7 (16.7)	40.9 (16.9)	44.4 (18.6)	-0.224	0.017
n_sec	20.8 (7.8)	19.2 (7.1)	20.9 (7.9)	-0.213	-0.018
log_text_length	4.0 (0.2)	4.0 (0.2)	4.0 (0.2)	-0.355	-0.002
text_ppl	0.8 (0.0)	0.8 (0.0)	0.8 (0.0)	0.172	-0.015
topic_cluster				0.188	<0.001
00	RL/DL/Robustness	325 (4.3)	77 (5.2)		
01	RL/DL/CV	66 (4.4)	392 (5.2)		
02	DL/Generative Models/CNN	46 (3.1)	223 (3.0)		
03	DL/RNN/GNN	55 (3.7)	340 (4.5)		
04	DL/Optimization/Generalization	140 (9.4)	495 (6.6)		
05	DL/Robustness/Adversarial Examples	106 (7.1)	458 (6.1)		
06	DL/RNN/GNN	87 (5.9)	363 (4.8)		
07	RL/Multi-Agent RL/DL	129 (8.7)	752 (10.0)		
08	Federated Learning/DL	62 (4.2)	271 (3.6)		
09	Generative Models/VAE/GAN	74 (5.0)	465 (6.2)		
10	DL/NLP/Transformer/LM	108 (7.3)	543 (7.2)		
11	GNN/GCNN/Representation Learning	94 (6.3)	387 (5.2)		
12	DL/Self-Supervised Learning/Meta-Learning	87 (5.9)	455 (6.1)		
13	RL/GNN/Transformer	49 (3.3)	289 (3.9)		
14	DL/Model Compression/Neural Architecture Search	48 (3.2)	332 (4.4)		
15	DL/Representation Learning/Transfer Learning	44 (3.0)	274 (3.7)		
16	DL/Representation Learning/Word Embeddings/NLP	41 (2.8)	251 (3.3)		
17	DL/Neural Architecture Search/Optimization	50 (3.4)	296 (4.0)		
18	DL/RL/Representation Learning/Generalization	57 (3.8)	299 (4.0)		
19	DL/Interpretability/Uncertainty Estimation	66 (4.4)	283 (3.8)		
n_author	4.2 (1.9)	4.3 (2.0)	4.1 (1.9)	0.096	0.002
n_author_female	0.3 (0.6)	0.4 (0.6)	0.3 (0.6)	0.087	0.009
first_author_female				0.106	0.006
True	87 (5.9)	643 (8.6)	89 (6.0)		
False	1399 (94.1)	6850 (91.4)	1397 (94.0)		
any_author_female				0.085	0.022
True	375 (25.2)	2173 (29.0)	361 (24.3)		
False	1111 (74.8)	5320 (71.0)	1125 (75.7)		
no_US_author				0.056	0.006
True	464 (31.2)	2537 (33.9)	468 (31.5)		
False	1022 (68.8)	4956 (66.1)	1018 (68.5)		
log_ins_rank_min	1.1 (0.7)	1.1 (0.7)	1.1 (0.7)	0.042	-0.001
log_ins_rank_avg	1.6 (0.6)	1.5 (0.6)	1.6 (0.6)	-0.041	-0.000
log_ins_rank_max	1.8 (0.6)	1.7 (0.6)	1.8 (0.6)	-0.064	-0.001
log_author_cite_min	2.3 (0.8)	2.3 (0.8)	2.4 (0.8)	0.008	-0.021
log_author_cite_avg	3.6 (0.6)	3.6 (0.7)	3.6 (0.6)	-0.067	<0.001
log_author_cite_max	4.0 (0.7)	3.9 (0.8)	3.9 (0.7)	-0.083	0.012

Table A.1: **Summary of covariates before and after matching.** We show the mean and standard deviation (in parentheses) for each covariate, and the standardized mean difference (SMD) between the treated group and unmatched and matched comparison groups. Note that after matching the SMD reduces significantly for all covariates and a fine-balance has been achieved in the topic_cluster. In topic_cluster we also display top keywords associated with each cluster. Abbreviations: RL = Reinforcement Learning, DL = Deep Learning, CV = Computer Vision, [GC]NN = [Graph Convolutional] Neural Nets, VAE = Variational Autoencoder, GAN = Generative Adversarial Network.

Appendix A. Data Preparation

We build on data collected and processed by Zhang et al. (2022) using the ICLR data between the years 2018–2022.¹³ The original dataset includes 10,297 submissions between 2017–2022, consisting of all the submitted papers to ICLR between these years, not including retracted papers. Submissions in the year of 2017 are not used in the study as the treated group contains too few samples.

13. cogcomp.github.io/iclr_database

	Accepted	Rejected
# submissions	3,678	6,619
Has S2 ID	3,636 / 3,678 (99%)	5,336 / 6,619 (81%)
Has date	3,209 / 3,636 (88%)	4,879 / 5,336 (91%)

Table A.2: Statistics from the Semantic Scholar (S2) data processing. We present the total number of considered submissions (*# submissions*), the fraction of submissions we successfully matched to an S2 document (*Has S2 ID*), and the fraction of submissions with available publication date from the S2 matched submissions (*Has date*).

A.1. Confounders

We consider 18 confounders, summarized in Table A.1. We now describe the processing of these variables, divided into two high level categories: *submission metadata* and *submission content*.

Submission Metadata We use the following metadata information: submission year, number of authors, number of female-identified authors,¹⁴ whether the first author is identifying as female, and whether there is a non-US base author. In addition, we use some features that account for the authors institute, and citation count: the highest, lowest, and average institution ranking,¹⁵ and the highest, lowest, and average citation counts of the authors.¹⁶

Submission Content Using the information extracted from the papers, we take into account several content-related potential confounders. We consider the number of figures, references and sections of a paper extracted using Grobid¹⁷ on the fulltext. We also use the length of the paper (in log scale) measured using Longformer tokenizer (Beltagy et al., 2020), and a measure of paraphrase fluency¹⁸ averaged over the full document (calculated by averaging over paraphrased token likelihoods from the pre-trained RoBERTa model (Liu et al., 2019), normalized to (0, 1), where higher values indicate the utterances are more likely, For simplicity, we refer to this as `text_perplexity`). Finally, we consider 20 topic clusters, computed by spectral clustering (Ng et al., 2001) using the Sepcter embedding (Cohan et al., 2020) of abstracts in scikit-learn (Pedregosa et al., 2011).

A.2. Paper Citations

In Section 4, the citation count $CC^{(n)}$ is described as the number of times that a given paper was cited in the n -year window after it was first made public. As a concrete example, for a paper that first appeared online as an arXiv preprint on May 15, 2018, $CC^{(2)}$ counts the number of citations received by this paper as of May 15, 2020. For a paper that was never arXived and first appeared online as part of conference proceedings published on October 20, 2018, $CC^{(2)}$ counts the number of citations received by this paper as of October 20, 2020. This allows for an “apples-to-apples” comparison of papers’ citation counts; by contrast, comparing citation counts by (for instance) calendar year

14. The gender is based on author’s OpenReview profile. We construct the indicator variable of whether an author is self reported to be female to be used in our analysis.

15. Institute ranking is determined by counting the total number of papers accepted to ICLR in previous years.

16. Author citation counts are obtained from the Google Scholar API (Cholewiak et al., 2021), as of February, 2022.

17. github.com/kermitt2/grobid

18. github.com/PrithivirajDamodaran/Parrot_Paraphraser

NCO	$CC^{(1)}$	$CC^{(2)}$	$CC^{(3)}$
Conference Years	2018 – 2022	2018 – 2021	2018 – 2020
Matched Pairs	1,486	1,073	570

Table B.1: **Number of Matched Pairs.** The covered years of the ICLR conference, together with the matched pairs found in the data for each negative control outcome variable ($CC^{(n)}$) decision (the number of citations after $n \in \{1, 2, 3\}$ years).

would inflate the citation counts of papers preprinted on arXiv prior to their appearance in conference proceedings, relative to papers that did not first appear as preprints.

We describe our process of obtaining $CC^{(n)}$ using the Semantic Scholar Academic Graph (Kinney et al., 2023).

Matching ICLR papers to Semantic Scholar First, we use the Semantic Scholar (S2) API to match each submitted ICLR paper to a unique document in S2; this is done using a fuzzy match on paper title. The S2 pipeline includes a canonicalization step where multiple instances of the same paper from different sources are merged into a single entry; for instance, a paper that was released first on arXiv and then included in the ICLR conference proceedings is represented as a single paper in S2, and its citations are merged (i.e. citations of both the arXiv preprint and the conference article are included in the canonical citation count).

Next, we query the S2 Academic Graph API for metadata on each matched paper, including its publication date. For S2, the publication date is not the day of the conference at which the paper was presented, but rather the *date that the paper first became available through any of the data sources S2 ingests* (including arXiv, DBLP, the ACL anthology, etc.). Effectively, it is the first day when the paper became visible online.

Table A.2 shows the results of these data processing steps. We were able to match nearly all accepted ICLR papers to a paper in S2. For rejected papers, we matched 81%. The remaining 19% of rejected papers were likely never posted to arXiv or published elsewhere. In our analysis, we assign these papers 0 citations since they were never published and hence never cited. Of the papers where we found a matching paper in S2, roughly 90% of both accepted and rejected papers had a publication date available in their metadata; the 10% with no publication dates were removed. Publication dates were missing in 6% of the early arXived papers, and 11% of the rest.

Counting Paper Citations For each paper P with an available publication date, we queried the S2 Academic Graph API for all papers citing P , and submitted a followup query requesting the publication dates for all citing papers. Roughly 10% of citing papers did not have publication dates available; these were discarded. With the publication date of P and all its citing papers in hand, we compute $CC^{(n)}$ by counting the number of citing papers which were published (i.e. became available online) within n years of the publication of P .

Appendix B. Setup Details

Matched Pairs We perform tripartite matching (Zhang et al., 2021) under the same setup as in Chen et al. (2022) such that the (1) discrepancy between numerical covariates is minimal; (2) the categorical variables `n_author` and `year` are nearly exactly matched; (3) the distribution of

Year	Early arXiving (A)	# Accept (Total Cnt)	# Reject (Total Cnt)	CC ⁽¹⁾ (Average)	CC ⁽²⁾ (Average)	CC ⁽³⁾ (Average)
2018	A = 0	10	10	9.85	27.45	47.40
	A = 1	19	1	30.10	112.60	249.95
2019	A = 0	24	24	14.52	38.70	63.52
	A = 1	27	21	14.13	39.04	69.79
2020	A = 0	190	312	8.32	20.53	34.03
	A = 1	235	267	11.86	34.50	58.14
2021	A = 0	182	320	8.99	25.15	-
	A = 1	232	271	8.66	24.33	-
2022	A = 0	198	216	8.76	-	-
	A = 1	236	177	7.81	-	-

Table B.2: **Summary of paper decision and n -year citation counts (average) in the matched sample.** Although the number of accepted papers are larger in the early arXived groups, long-term citation counts are also higher. This suggests that it is likely an unobserved confounder (paper quality) that affect both the paper acceptance and the long-term citation count.

Author Institution Early arXiving (A)	CC ⁽ⁿ⁾ Qtl	Top-10		Top-10 to 100		Others	
		A = 0	A = 1	A = 0	A = 1	A = 0	A = 1
CC ⁽¹⁾	50% (4)	36	46	320	398	214	322
	75% (11)	14	16	171	225	86	139
	90% (23)	5	4	77	102	26	42
CC ⁽²⁾	50% (10)	26	39	247	304	141	232
	75% (29)	13	16	133	172	54	108
	90% (65)	3	6	52	85	16	34
CC ⁽³⁾	50% (11)	20	26	168	205	91	157
	75% (41)	8	11	95	118	32	75
	90% (103)	2	4	34	59	9	28

Table B.3: **Sample sizes in each subgroup when stratified by author institution ranking.**

topic_cluster in treated and matched groups are similar (also known as *fine-balance*). Recall that when performing analysis using NCOs (CC⁽ⁿ⁾), only a subset of data with conference year in or before the year (2023 - n) is used. We tabulate in Table B.1 the sample size (as the number of matched pairs) for each subset.

Motivation for Citation Counts As shown in Table B.2, although early arXived papers have higher rate of acceptance in the matched sample, long-term citation counts are also higher. This indicates that it is likely that an unobserved confounder (e.g., paper quality) affects both of them.

Stratification When we perform stratified analysis, the same strata may contain a different number of samples depending on the citation counts we use. The breakdown of sample sizes is tabulated in Table B.3 for stratification under author institution ranking, and in Table B.4 for stratification under author citation counts. We also tabulate the thresholds corresponding to each quantile in these tables. We attempt to find reasonable thresholds for binning the strata while ensuring a non-zero sample size in each of them, and covariates related to the subgroup stratification are discarded (institute_rank

Author Citation		<500		500-2000		>2000	
Early arXiving (A)		A = 0	A = 1	A = 0	A = 1	A = 0	A = 1
CC ⁽ⁿ⁾	Qtl						
CC ⁽¹⁾	50% (4)	16	25	54	64	500	677
	75% (11)	10	9	15	25	246	346
	90% (23)	2	2	7	7	99	139
CC ⁽²⁾	50% (10)	15	15	40	52	359	508
	75% (29)	9	7	15	21	176	268
	90% (65)	2	2	4	6	65	117
CC ⁽³⁾	50% (11)	15	16	27	32	237	340
	75% (41)	9	8	9	16	117	180
	90% (103)	2	2	2	6	41	83

Table B.4: Sample sizes in each subgroup when stratified by author citation count.

and `author_citations` respectively). We acknowledge that some strata have few samples, making it hard for statistical procedures to be sensible, which renders a considerably wider confidence interval for estimated effect on the strata.

Appendix C. Causal Inference: More Technical Details

C.1. Relaxing the DiD Assumption: Quantile-Quantile Equi-Confounding

In this section we explore an alternative to the *additive equi-confounding* assumption, the core assumption in using NCO through DiD to deconfound the effect. Our motivation is to eliminate two undesirable consequences that the additive equi-confounding assumption introduced:

1. N and Y are assumed to be affected by the *same* unobserved confounders; and
2. unobserved confounders are assumed to affect N and Y at a similar scale.

In the analyses presented in the main text, we assume (1) and operationalize (2) by dichotomizing N (thus N is scaled to a similar level as Y). As an alternative, [Sofer et al. \(2016\)](#) proposed the *quantile-quantile equi-confounding* (QQ equi-confounding for short) assumption (for interested readers, this is an analogue in NCO of the more well-known change-in-change method studied in [Athey and Imbens \(2006\)](#)), which is weaker than the additive equi-confounding assumption. QQ equi-confounding works when Y and N might be affected by a possibly different set of unobserved confounders, and is invariant to the scaling of N . In this sense, both (1) and (2) above can be relaxed in QQ equi-confounding.

We first collect some definitions and results from [Sofer et al. \(2016\)](#). Write $F_{X|Z}(x)$ as the cumulative distribution function (CDF) of X given Z , and define $F_{X|Z}^{-1}$ as its inverse function. Assume for now we write U, W for the unobserved confounder for Y and N respectively (U and W may also be associated with A and between themselves), the quantile-quantile (QQ) association between U and A conditional on C , and between W and A conditional on C are defined as:

$$q_0(u | c) = F_{U|A=0,C=c} \circ F_{U|A=1,C=c}^{-1}(u),$$

$$q_1(w | c) = F_{W|A=0,C=c} \circ F_{W|A=1,C=c}^{-1}(w),$$

for $u, w \in [0, 1]$. The alternative QQ equi-confounding assumption is then framed as follows.

Assumption C.1 (QQ Equi-confounding)

$$q_0(v | c) = q_1(v | c), \quad v \in [0, 1].$$

This assumptions means, after transformed into the quantile level, the association between U and N is the same as between W and Y , conditional on C . Note that since quantiles are invariant under monotone transforms, this assumption does not restrict the scale between Y and N to be the same.

Under the framework of QQ equi-confounding, an alternative positivity assumptions is needed. Let $N^* \sim (N | A = 1, C)$ be a random variable distributed the same as the NCO in the treated group, positivity assumes that if $0 < f_{N|A=1,C}(N^*)$, then $0 < f_{N|A=0,C}(N^*) < 1$, where f is the probability density function corresponding to F . This assumption is the analogue of positivity on the quantile level, which we assume.

Theorem 1 (ATET Under QQ Equi-Confounding (Theorem 1 from Sofer et al. 2016)) *Under the above assumptions, ATET can be expressed as:*

$$\text{ATET} = \mathbb{E}[Y | A = 1] - \mathbb{E}[\tilde{Y}],$$

where

$$\tilde{Y} = F_{Y|A=0,C}^{-1} \circ F_{N|A=0,C}(N^*).$$

As a first step, we examine the QQ transformation,

$$\text{qq}(u) = \mathbb{E}_C \left[\hat{F}_{Y|A=0,C}^{-1}(\hat{F}_{N|A=0,C}^{-1}(u)) \right]$$

as a function of $u \in [0, 1]$, where the expectation is taken with respect to covariates distributed as in the treated group. We observe in Figure C.1 that the QQ plot departs from the identity (dashed) lines, which encodes unobserved confounding (Sofer et al., 2016). Using $N = \text{CC}^{(3)}$ as the NCO, under the QQ equi-confounding assumption, we estimate that

$$\text{ATET} = -4.375\%,$$

with 95%-bootstrap confidence interval being

$$(-9.965\%, -0.092\%).$$

Discussion This confidence interval almost touches 0, indicating that under the QQ equi-confounding assumptions, estimated effects are very weak. Comparing the results with our analysis in Figure 2, though the direction of effect may differ, the significant levels are consistent: in both analyses we do not have significant (or very weak) evidence for early arXiving to affect paper decision (either positively or negatively).

Although under the QQ equi-confounding setup, we are able to circumvent the two limitations of the additive equi-confounding outlined at the beginning of this section, our choice of using the citation count $\text{CC}^{(n)}$ might not be the best: as discussed in Section 1, the citation count itself may not be the best NCO since early arXiving might as well causally affects how it varies at a minor scale (e.g., to change it from a value of 10 to 11, though both may be view as “less-cited”). A more elaborated study would entail “smoothing out” minor variations in long-term citation counts while preserving as much information as it could. Such explorations are out of the scope of this appendix and we welcome future work along this direction.

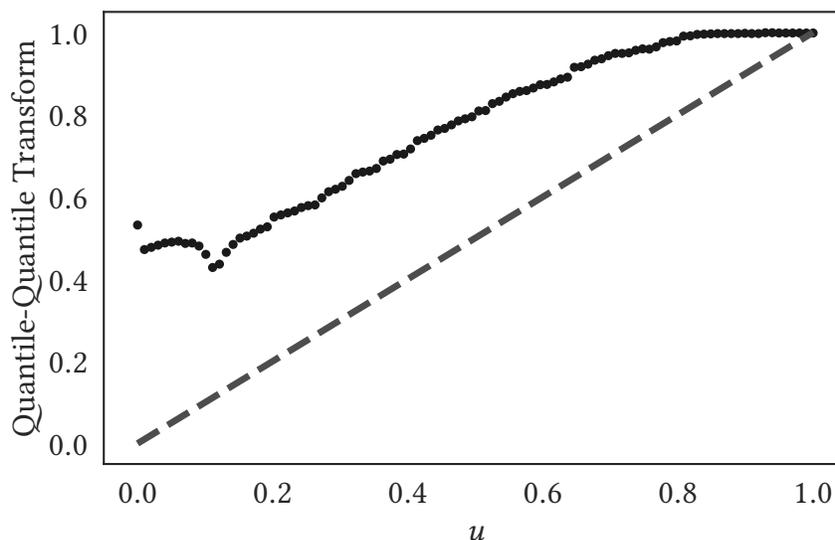


Figure C.1: **Empirical QQ-Plot** ($qq(u)$). The departure from the identity (dashed line) encodes unobserved confounding.

Appendix D. Dichotomize or Not?

In our main analysis using DiD for NCO, we dichotomize the citation counts $CC^{(n)}$ at various thresholds $CC_q^{(n)}$ defined by the empirical quantiles in the matched sample. There are two motivations for dichotomization.

1. Dichotomization made Y and N have the same scale, as required by the DiD additive equi-confounding assumption.
2. As we view “highly cited” from long-term citation counts as an NCO, fluctuations at a minor scale (e.g., cited by 10 times vs. 11 times) should be discarded. Dichotomization naturally “smooth” out these fluctuations.

The limitation of dichotomization is also obvious: (1) we do not know what is the “correct” threshold for dichotomizing; (2) much information was lost during this process. To provide a comprehensive study, for each citation count $CC^{(n)}$, we perform three studies using three NCOs by thresholding $CC^{(n)}$ at three levels. The results are interpreted under the assumption of dichotomizing at the level is valid, which is uncheckable. On the other hand, to provide another perspective when all information is kept, in Appendix C.1, we explore an alternative methods in which $CC^{(n)}$ is directly used as an NCO without dichotomizing.