HOW IMPORTANT IS OPEN-SOURCE SCIENCE FOR INVENTION SPEED

Tan Tran

Evens Salies

SCIENCES PO OFCE WORKING PAPER n° 10/2023

EDITORIAL BOARD

CONTACT US

OFCE
10 place de Catalogne | 75014 Paris | France
Tél. +33 1 44 18 54 24
www.ofce.fr

WORKING PAPER CITATION

This Working Paper:
Tan Tran and Evens Salies
How important is open-source science for invention speed
Sciences Po OFCE Working Paper, n° 10/2023.
Downloaded from URL: www.ofce.sciences-po.fr/pdf/dtravail/WP2023-10.pdf
DOI - ISSN

ABOUT THE AUTHORS

Tan Tran, ICD business School, IGS-Groupe
Email Address: tttran@groupe-igs.fr
Evens Salies, OFCE-Sciences Po
Email Address: evens.salies@sciencespo.fr

ABSTRACT

This study questions whether not citing open-source science lowers invention speed. We use a unique observational data on clinical trials projects submitted during the early months of Covid-19 spread. To estimate the effect of open-source science on invention speed we employ matching methods combined with regression. Our results show that, on average, projects that use open-source science can be accelerated by 51 days. We also estimate the effect of open-source science within the subsample of control projects. The effect, however, is less than that for projects that use open-source science.

KEYWORDS

Treatment effects, Variables selection algorithm, Open-source science.

JEL

C21, C52, O36.

RÉSUMÉ

Cette étude se penche sur la question de savoir si la citation de travaux en accès libre accélère le processus d'invention. Nous utilisons des données observationnelles de projets d'essais cliniques soumis au cours des premiers mois de la propagation de la Covid-19. Pour évaluer l'impact de la science en accès libre sur la vitesse d'invention, nous employons des méthodes d'appariement et de régression combinées. Nos résultats montrent qu'en moyenne, les projets qui intègrent des références à des travaux en accès libre sont soumis 51 jours plus tôt. L'impact de la disponibilité de la science en accès libre dans l'échantillon de projets témoins est moindre que celui observé dans les projets qui utilisent la science en accès libre.

MOTS CLÉS

Inférence causale, Algorithme de sélection de variables, Science ouverte.

JEL

C21, C52, O36.

HOW IMPORTANT IS OPEN-SOURCE SCIENCE FOR INVENTION SPEED

Tan Tran[1], Evens Salies[2,3]

[1] ICD business School, IGS-Groupe, 12 Rue Alexandre Parodi, 75010 Paris, France. E-mail: tttran@groupe-igs.fr (corresponding author)

[2] Sciences Po, OFCE, Paris, France. E-mail : evens.salies@sciencespo.fr

Abstract

This study questions whether not citing open-source science lowers invention speed. We use a unique observational data on clinical trials projects submitted during the early months of Covid-19 spread. To estimate the effect of open-source science on invention speed we employ matching methods combined with regression. Our results show that, on average, projects that use open-source science can be accelerated by 51 days. We also estimate the effect of open-source science within the subsample of control projects. The effect, however, is less than that for projects that use open-source science.

Keywords: Treatment effects, Variables selection algorithm, Open-source science
JEL Codes: C21, C52, O36

Résumé

Cette étude se penche sur la question de savoir si la citation de travaux en accès libre accélère le processus d'invention. Nous utilisons des données observationnelles de projets d'essais cliniques soumis au cours des premiers mois de la propagation de la Covid-19. Pour évaluer l'impact de la science en accès libre sur la vitesse d'invention, nous employons des méthodes d'appariement et de régression combinées. Nos résultats montrent qu'en moyenne, les projets qui intègrent des références à des travaux en accès libre sont soumis 51 jours plus tôt. L'impact de la disponibilité de la science en accès libre dans l'échantillon de projets témoins est moindre que celui observé dans les projets qui utilisent la science en accès libre.

Mots clés: Inférence causale, Algorithme de sélection de variables, Science ouverte
Classification JEL: C21, C52, O36

[3] **Acknowledgement**. This paper has benefited from helpful suggestions by two discussants at the 4th International Conference on Digital, Innovation, Financing and Entrepreneurship, Concordia University, Canada. We also thank Kate Paul for assistance in finalizing the text of this paper. Any remaining errors are our entire responsibility.

1. Introduction

Market competitiveness, technological progress, and customer choice, as well as regulatory restrictiveness create the need for faster product development. This requires firms to couple external factors with organizational capabilities for speed.

Kessler and Chakrabarti (1996, p. 1143) define innovation speed as "the time elapsed between initial development and ultimate commercialization, which is the introduction of a new product into the marketplace". Invention is distinguished from innovation by "a time lag and a continuous development process that fills this time gap and turns knowledge into a product" (Milan et al., 2020, p. 4). This process combines a variety of types of knowledge, capabilities, skills, and resources (Fagerberg et al., 2005). Invention speed concerns the fast process from initial idea, to search, and develop product, not yet in commercialization stage; see Stalk and Hout (1990) for more details.

Among external factors, Open science, which involves sharing of knowledge and ideas without proprietary protection, has been suggested as one possible key determinant of innovation speed. Open science emerged in late 16th-early 17th century, which placed Europe as the global innovative civilization for centuries (David, 2008). Over the past decades, policy changes made the imbalance between open science and intellectual proprietary. Open science aims to produce new knowledge while the proprietary model, which increases transaction costs and reduces knowledge diffusion through patent tickets and licensing (Galasso and Schankerman, 2015), captures value through commercializing innovation (Dasgupta and David, 1994).

Open science (access to scientific outputs and processes, such as open publications, data, or reproducibility of results and open peer review, open-source software) has been re-discussed as a mechanism to fix the failure of the innovation system for priviate propriety and socical benefits. Particularly, to deal with the covid-19 pandemic, an open science partnership between firms and researchers quickly share data, ideas and tools to accelerate research and innovation (see Gold, 2021). Also, Magalhaes et al. (2017) emphasize the role of open science in times of big data and innovation. In this new era open science creates an information network and collaborative intelligence that help to fight and to decimate the health ills plauging our society. Open science reduces the sense of ownership that researchers have over their data. It also increases equity and economic return of public investment since the data are often the results of research funded with public money and materials donated by others. These authors, however, do not explicitly differentiate open and closed sources. It is expected that they have likely different effects on innovation speed. Consequently, the lack of study on the effect of OSS

prevents scholars to fully understand its role on inventive activity. Our main objective in this paper is to fill this gap by identifying research projects that rely on open- and closed-source science.

The present paper focusses on invention rather than innovation speed and its relation to OSS. We investigate whether researchers citing OSS invent faster than those not doing so. We focus on a particular period where new research findings could have worldwide consequences. More specifically, we use unique observational data on clinical trials projects submitted in the early months of the Covid-19 pandemic until December 2021.

The members of the teams' decisions to use or not OSS in the past is necessarily confounded given the observational nature of our data. Plausible confounders are the total amount of funding received by projects' members in the past and the number of data sets released by members in the project to mention a few. These and other pre-treatment variables will be used to mitigate selection bias at the project level. We assess covariate balance using standard univariate and multivariate criteria. The propensity score uses variables selected by using Imbens and Rubin's (2015) algorithm which we implemented in Stata.[4]

We follow the potential outcomes approach to causal inference for binary treatments: see Imbens and Rubin (2015). A crucial step in that framework is estimating what would have been invention speed for projects that rely on OSS had they not relied on OSS. We estimate this unobserved outcome and the effect of OSS on invention speed by using regression as a benchmark, then we use the more robust bias-corrected matching estimator of Abadie and Imbens (2011).

Our results show that using OSS increases speed by 51 days. We also find that control projects would have invented solutions earlier, had they used OSS. The gain, however, is slightly lower than that of citing OSS. Our results have policy and practical implications. First, alongside other factors accelerating invention (funds and online platforms), OSS can be encouraged by the public sector as an efficient complementary tool to intellectual proprietary. Second, the private sector would benefit in having more firms sharing their invention.

The plan is as follows. In section 2 we address why OSS can be used as a mechanism for speeding up invention in normal times and crisis times such as during the Covid-19. Section 3 describes our data. Section 4 introduces the econometric model and matching estimation strategy. Results are given in section 5, before policy recommendation and conclusion in section 6.

[4] Code and data are **freely available** from the authors on request.

2. Why OSS is important for invention speed?

The Covid-19 pandemic is an example under which OSS was expected to alleviate restricted mobility and time constraint. The situation was paradoxical: scientists were urged to accelerate their invention while dealing with a loss of personal contacts with the scientific community. Relying on the advice of one's colleagues within the lab was limited. Moreover, to identify external expertise who could help was an even more challenging task during the pandemic. In addition, Noble and Spanjol (2020) highlighted that data collection was more challenging. They emphasize the importance of open publishing platforms. Instead, scientists could access scientific outputs, more particularly OSS.

During the pandemic, the White House Office of Science and Technology Policy openly announced all relevant research on the Covid-19 to encourage scientists to collectively respond to the crisis and to work on solutions. They also called to action to the tech community to release the most extensive machine-readable Coronavirus literature,[5] including data and full text of scientific articles. The main goal was to spread information of gene sequencing of the virus and to share medical resources, databases as quickly as possible to boost discoveries, testing, and approval of potential solutions to the covid-19 pandemic. This initiative allows scientists to develop treatments, vaccines, and drugs not only for now, but also for future pandemics.

Chesbrough (2020) showed how firms responded to the Covid-19 by comparing their incentive for using open innovation during the Covid-19 with those motivations in normal times. The author suggested that openness helped to mobilize knowledge from different places, boosted our learning to advance and our progress against the covid-19 disease to accelerate. Openness also encouraged voluntary researchers who could utilize their own existing facilities from different countries and time zones. Furthermore, it leveraged both the human and physical capitals (plants and equipment) in the world to launch rapid solutions. We propose that OSS works under two mechanisms.

First, OSS generates potential coalitions whereby knowledge (data, publications, materials, and tools) is pooled into public platforms. This process removes barriers on sharing explicit and tacit knowledge and ensures relationships of trust of peer reviewers (Arthur, 2007). On one hand, the results of open science are freely available on the Internet[6]. Potential participants can

[5] https://trumpwhitehouse.archives.gov/briefings-statements/call-action-tech-community-new-machine-readable-covid-19-dataset/

[6] The peer-reviewed journals accept work that has previously been posted public.

briefly summarize their research program, hence reduce duplication of efforts. On the other hand, the free access to scientific knowledge (e.g., data and artifacts in the biomedical sector) motivates scientists to reengage in breakthrough research, explore novel hypotheses and reduce innovation costs. This lowers overall risk (Munos and Chin, 2011) and accelerate the decision making of teams.

Moreover, unlike closed-source science, OSS encourages culture of scientific knowledge sharing to a wider community, such as NGOs, universities, research institutions, firms, policymakers, and individuals. It helps to alleviate transaction costs involved in search for newly scientific knowledge useful to the ongoing project (Woelfle et al., 2011). The reduced costs make the project more viable to stakeholders (small firms, organizations, and users) and encourage them to join in the partnerships (see Gold, 2021 on the participation of outside experts). The research program inevitably accelerates faster than if scientists had attempted to reach others within their limited professional network individually (Woelfle et al., 2011). There is a case in point in the drug industry: the praziquantel (PZQ), which is used in the treatment of schistosomiasis infection (Hotez et al., 2010). In January 2010, all data of preliminary experiments were published in an open-source platform. Though promising, some research projects raised problems that they were unable to address. In April, a request for suggestions was posted on LinkedIn. The PZQ project team received comments from unknown experts. By mid-May, the team decided to send a small sample of racemic PZQamine to one of those people, a Dutch contract research organization. On 25-Aug this latter posted a solution.

To sum up, OSS can be used as a mechanism for speeding up invention, not only in normal times but also in crisis times.

3. Data collection

We have two data bases, at the clinical trial level (*ClinicalTrials*)[7] and researcher level (*Dimensions*).[8] We start from 6110 clinical projects (until 31-Dec-2021). We identify the name of the investigator(s) who participated into a clinical trial. Then we merge the name with the

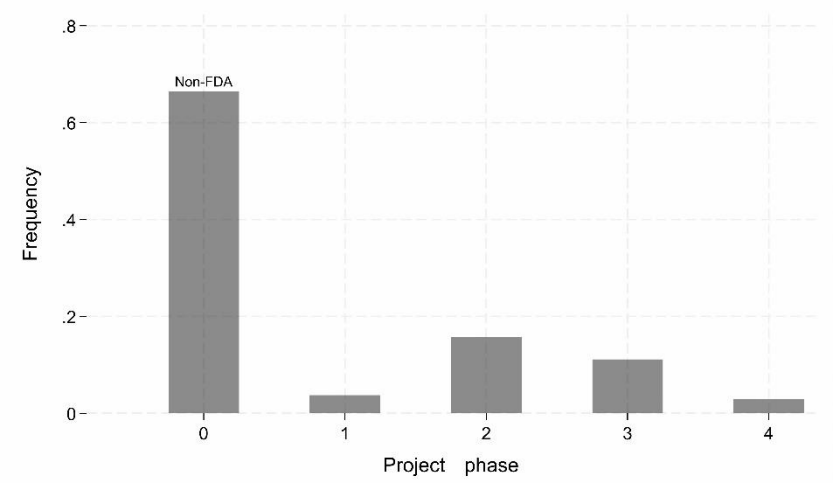
[7] This is a web-based resource that provides information of clinical studies on a wide range of diseases and conditions. National Library of Medicine at National Institutes of Health maintains the website.

[8] The *Dimensions* data base gathers all research information (papers, books, chapters, awarded grants, datasets and clinical trials, patents, and policy documents). This platform is harvested from sources such as CrossRef, PubMed, Directory of Open Access Journals, Open Citation Data, clinical trial registries, patent offices and over 100 publishers; URL: https://www.dimensions.ai/why-dimensions/ or https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6579599/.

researcher level data. This latter includes 5368 observations of scientists who registered on *ClinicalTrials* from the beginning of the Covid-19 outbreak period[9] to the last date of our data collection. The merged data includes 2433 clinical trial projects. Our variables are listed below. They include several potential confounders.

We have a variable that indicates at which stage the clinical trial is. In general, there are two types of *project phases*, namely FDA-defined and non-FDA-defined phases. The following graph shows the distribution of the *phase* variable across projects. The FDA-defined phase trial includes four main phases (Phase I, II, III and IV) of trials with patients, where Phase IV indicates projects that have been approved by FDA for use in the general population. More specifically, Phase I describes clinical trials that focus on the safety of a drug to be conducted with healthy volunteers. Phase II gathers preliminary data on whether a drug works in people who have a certain condition/disease. Phase III studies different populations and different dosages and by using the drug in combination with another drug. Phase IV is about gathering additional information about a drug's safety, efficacy, or optimal use. Phase Not Applicable is a trial without FDA-define phase (such as trials of devices or behavioral interventions).

Figure 1: Distribution of project phases



[9] The Economic Times, "The first COVID-19 case can be traced back to November 17 in China's Hubei province", March 13, 2020.

All clinical trials in our sample received funds, yet the specific amounts are not available. We must content ourselves with the total amount of funds received by researchers up to the date at which we collected the data. We call this variable *Past funding*. This amount is recorded in the *Dimensions* database. We aggregated the amount of funds at the project team level and took the logarithmic transformation. Furthermore, we created a dummy variable which takes the value 1 when none of the project members received any funds in the past. This variable will be used for exact matching to select comparison projects.

We control for scientific knowledge spilled over by team members to other teams by counting the number of datasets released by project members. It is the total number of open datasets that the team's members have openly published. We built an *outward-science* variable, which is 100 times the ratio of open datasets released by the team members to the maximum value of open datasets.

To capture the potential effect of team research ability to accelerate inventive discovery, we use the number of *publications* of all members (the variable sums up all previous published articles by project's team members), the number of *citations* received by team members aggregated at the project level (Heirman and Clarvsee, 2007). We also count total number of *patents* invented, and *trials* conducted by project's team members in the past. In addition, we observe the total number of distinct *co-authors* in publications of each team member in the past. We calculate this variable at the project level.

For each project we also use the number of *collaborators* to proxy for the diverse knowledge of team members (Bercovitz and Feldman, 2011). In normal research periods, researchers may take time to form a team. However, mobility restrictions during the Covid-19 crisis forced team members to select partners working effectively (Khanagha et al., 2021).

The outcome variable, *speed of invention*, is used to measure the extent to which a project is ahead of the others. The measure is calculated in two steps. First, we measure the total number of days that have elapsed since the outbreak date. For each clinical trial, it is a function of the number of days passed between the date of the Covid-19 outbreak date (t_0) and the starting date of the clinical trial i, say t_i. We followed previous studies to transform the time lag value $t_i - t_0$ (see Feldman et al., 2023). Between the outbreak date and the latest date (the date at which we collected the data) in our sample, 775 days passed. *Speed* is defined as $775 - (t_i - t_0)$. The minimum value of *Speed* is 30, which presents the greatest time lag. The maximum value is 761, corresponding to a time lag value of 14 days. This is also the first clinical trial registered.

4. Econometric model

Our treatment variable for project i, OSS_i, takes the value 1 when the number of cited open-source articles is positive; it takes the value 0 otherwise. The speed of invention of project i can take two values, $Y_i(1)$ if the project uses OSS and $Y_i(0)$ otherwise.

Our quantity of interest is the average treatment effect on the treated $E(Y_i(1) - Y_i(0)|OSS_i = 1) \equiv ATT$. We will also measure the effect of OSS on invention speed within the subsample of control projects (ATU).

The combination of regression adjustment with matching is superior than using matching or regression separately: see Imbens and Rubin (2015, p. 432). To obtain ATT and ATU we rely on Abadie, Drukker, Herr, and Imbens (2004) implementation in Stata of Abadie and Imbens' (2011) bias-corrected nearest-neighbor matching estimator. This estimator allows each unit to be used as a match more than once.

Let us denote the full set of observed covariates by $\boldsymbol{X}_i$ and the probability of using OSS by $e(\boldsymbol{x}) \equiv \Pr(OSS_i = 1|\boldsymbol{X}_i = \boldsymbol{x})$. At least two crucial assumptions underly the use of this estimator: assignment to treatment is probabilistic ($0 < e(\boldsymbol{x}) < 1$) and unconfounded that is $(Y_i(0), Y_i(1)) \perp OSS_i|\boldsymbol{X}_i$, where '$\perp$' means independence in distribution. Given these assumptions and the balancing property of $e(\boldsymbol{X}_i)$, assignment to treatment is also unconfounded given $e(\boldsymbol{X}_i)$. This result from Rosenbaum and Rubin (1983) can be used to assess multivariate balance.

We also make the "stability" assumption that using OSS in one project does not affect speed for other projects; interaction between team members and peer effects, in response to OSS, remain essentially within the team. Formally, the value of D_j for $j \neq i$ has no effect on how i responds to D_i: $Y_i(OSS_i, \boldsymbol{OSS}_{-i}) = Y_i(OSS_i)$, $\forall i$, where $\boldsymbol{OSS}_{-i}$ denotes the vector of treatments for all projects $j \neq i$. Obviously, this assumption is more plausible for projects that were invented earlier. More assumptions are given in Abadie and Imbens' (2011) paper such as the Lipschitz nature of $E(Y_i(d)|\boldsymbol{X}_i = \boldsymbol{x})$, $\forall\, d \in \{0,1\}$.

Under these assumptions, an estimator of $E(Y_i(0)|OSS_i = 1)$ is $E(E(Y_i|OSS_i = 0, \boldsymbol{X}_i)|OSS_i = 1)$ where $E(Y_i|OSS_i, \boldsymbol{X}_i)$ are average speeds in matched samples.

The bias in $\widehat{ATT}$ is due to inexact matching (see Imbens and Rubin, 2015, p. 416). The matching estimator with bias correction is given in Abadie and Imbens (2011, appendix). Its structure depends on the treatment effect we want to estimate. The estimator for ATT:

$$\tau_M^m = \frac{1}{N_1}\sum_{i:OSS_i=1}\left[Y_i - \frac{1}{M}\sum_{j\in J_{M(i)}}\left(Y_j + \hat{\mu}_0(\boldsymbol{X}_i) - \hat{\mu}_0(\boldsymbol{X}_j)\right)\right],$$

where N_1 is the number of clinical projects using OSS, M is the number of matches (we explore values ranging from 1 to 20 in the present paper), $J_M(i)$ is the set of clinical projects in the control group that are nearest-neighbor matches for i, and $\hat{\mu}_0(\boldsymbol{X}_i) - \hat{\mu}_0(\boldsymbol{X}_j)$ is the bias correction obtained, e.g. by regression. Matching without correction is not efficient when M is fixed. Under the above regularity conditions, the bias in $\widehat{ATT}$ is quickly attenuated provided one can find good matches.

5. Results

Let us denote the group means and unbiased variance estimators by $\bar{X}_d$ and $S_d^2, d = 0,1$, respectively. Table 1 shows $\bar{X}_d$ and S_d^2 of covariates used in this study. The last column includes Rosenbaum and Rubin's (1985) normalized difference $(\bar{X}_1 - \bar{X}_0)/((S_0^2 + S_1^2)/2)^{1/2}$ to assess the balance in covariate distributions.

Table 1. Summary statistics

	All ($N = 2433$)		Controls ($N_0 = 1963$)		Treated ($N_1 = 470$)		Norm. Diff.[b]
Pre-treatment var.	Mean	s.d.	Mean	s.d.	Mean	s.d.	
Phase	.80	1.22	.76	1.19	.97	1.31	.165
Outward-science	.04	.09	.04	.09	.04	.11	.022
Citations received	7508.01	16801.35	7310.07	16571.60	8334.74	17723.26	.059
Collaborators	2.21	3.60	2.20	3.62	2.21064	3.52	.001
Co-authors	857.45	1824.67	841.98	1846.40	922.03	1731.03	.044
Past funding (log) [a]	15.08	2.23	15.17	2.25	14.71	2.12	.210
Funding dummy	.61	.487	.61	.49	.61	.49	.001
Publications	248.19	461.70	241.42	451.35	276.45	502.17	.073
Patents	2.45	13.87	2.62	14.60	1.72	10.22	.071
Trials	6.05	12.89	6.17	13.31	5.52	10.98	.053
						Mahalanobis [c]	.317

Notes:

a. Average funding for projects with positive funding.

b. Absolute value of the normalized difference formula described in the text.

c. Mahalanobis distance $\left[(\bar{X}_1 - \bar{X}_0)'\hat{\Sigma}^{-1}(\bar{X}_1 - \bar{X}_0)\right]^{1/2}$ where $\hat{\Sigma}$ is the average of treatment groups' sample covariance matrices.

The differences between the two distributions are not large. There are two covariates (funding and phase) for which the means differ slightly between the two treatment groups. The difference between the means using the Mahalanobis distance is equal to .317, which is rather small, which is an indication of overall balance. To control for confounders, we specified a logit by using the interesting Imbens and Rubin's (2015) stepwise variable-selection algorithm.[10]

Table 2. Propensity score results (logit).[a]

Variable	Coefficient	z	$P > \|z\|$
Phase	.137	3.39	<.01
Outward science	−.524	−.76	>.40
Funding (log)	−.127	−3.24	<.01
Patents	−.008	−1.29	>.10
Funding dummy	−1.820	−3.14	<.01
Publications [b]	.807	4.11	<.01
Trials	−.025	−3.16	<.01
Constant	.279	.49	>.50
$N = 2433$			

Notes.

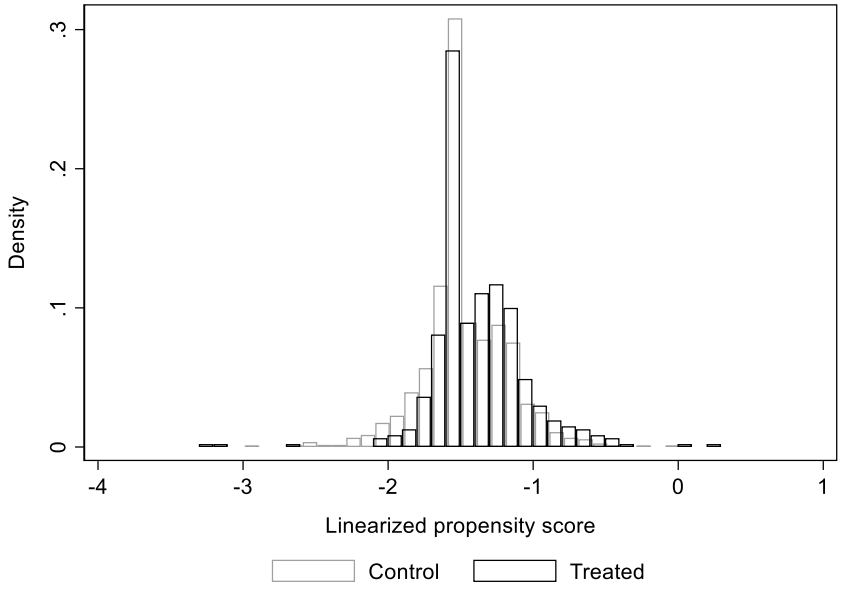
a. The variables have been selected by using the stepwise algorithm described in Imbens and Rubin (2015). The first three variables are imposed. No interaction and quadratic terms were selected by the algorithm.

b. The reported coefficient is 10^3 the actual coefficient.

Figure 2 below shows a histogram of the linearized propensity score (log-odds) for treated and control projects. The optimal number of blocks is six. The region of overlap is [.035, .548]; it is such that only six observations are removed from the analysis. The previous logit estimates are obtained in the overlap region.

[10] The code and data are available upon request.

Figure 2. Distribution of linearized propensity score for treated and control groups (in the overlap region)



The mean value of *speed of invention* in the whole sample is 471.4 days, and the standard deviation value is 153.2 days.

The results for the effect of OSS on speed obtained from different estimators are given in Table 3. Column 1 shows the estimate of the average treatment effect; it is 50.1 (s.e. $=$ 7.80, $p <$ 0.005),[11] which we obtained by simply regressing *speed of invention* on OSS. Although we ran the regression in the overlap region, this estimate is biased since we did not control for $\boldsymbol{X}_i$. When all the covariates are included in the regression function (these are the covariates selected by our implementation of Imbens and Rubin's algorithm), we obtain a close result, 49.9 (s.e. $=$ 7.54, $p < 0.005$), reflecting the good balance between most of the variables (column 2). The 'good balance' is an acceptable explanation since regression methods are not robust to substantial differences between treatment groups (Imbens, 2015).

To obtain the $\widehat{ATT}$ by using regression adjustment, we use the technique suggested by Rubin (1977), which consists in estimating $E(Y_i|OSS_i = 0, \boldsymbol{X}_i)$ in $E(E(Y_i|OSS_i = 0, \boldsymbol{X}_i)|OSS_i = 1)$ by regression; see Dehejia and Wahba (2002), Imbens (2015) and Imbens and Rubin (2015) for further details. We find 52.1 days ($p < 0.005$). The bootstrap standard error (s.e. $=$ 3.11) is lower than the previous ones.

[11] We follow recommendation of Benjamin, Berger et al. (2018) to set the threshold for defining statistical significance for new discoveries to less than 0.005.

Table 3. Treatment effect estimates of OSS on speed.

	ATE Simple OLS		ATE Multivariate OLS		ATT Matching by regression		ATT Bias-corrected NN matching ($M = 5$)	
	Coefficient	$P > \|z\|$	Coefficient	$P > \|z\|$	Coefficient	$P > \|z\|$	Coefficient	$P > \|z\|$
OSS_i	50.1	<.005	49.9	<.005	52.1	<.005	51.6	<.005
Phase			−6.9	<.05				
Outward science			29.5	>.18				
Funding (log) [a]			−8.9	<.005				
Patents			.3	>.33				
Funding dummy			−111.1	<.005				
Publications ($\times 10^{-3}$)			.8	>.49				
Trials			.5	>.21				
Constant	461.7	<.005	580.4	<.005				
$N = 2433$								

Note. The variables have been selected by using the stepwise algorithm described in Imbens and Rubin (2015). The first three variables are imposed. No interaction and quadratic terms were selected by the algorithm. All estimates are obtained in the overlap region with bootstrap standard errors (500 replications).

We then estimate ATT by using the bias-corrected nearest-neighbor matching (with replacement) estimator of Abadie and Imbens (2011). We select up to five matches from the group of control projects. We match groups exactly with respect to phase and the funding dummy. To measure the distance between the covariate distributions of open-source project i and potential matches, we use the Mahalanobis metric for five variables *funding*, *outward*, *publications*, *patents* and *trials*. We use our own weight matrix, namely the inverse of the average of the sample covariance matrix in the treatment group and the control group. Variables *collaborators*, *citations received* and *co_author* are used for the bias correction as described in the above paper. We find $\widehat{ATT} = 51.6$ ($p < 0.005$), which is lower than the value without correction not, 53.7, which we did not report in the table (53.7). This latter result suggests a positive bias (see Eq. 4 in Abadie and Imbens, 2011). These effects are slightly larger than the effect of OSS in the subsample of control projects ($\widehat{ATU} = 49.2$).

Figure 3. Distribution of τ_M^m for the ATT, $M = 1, \ldots, 20$

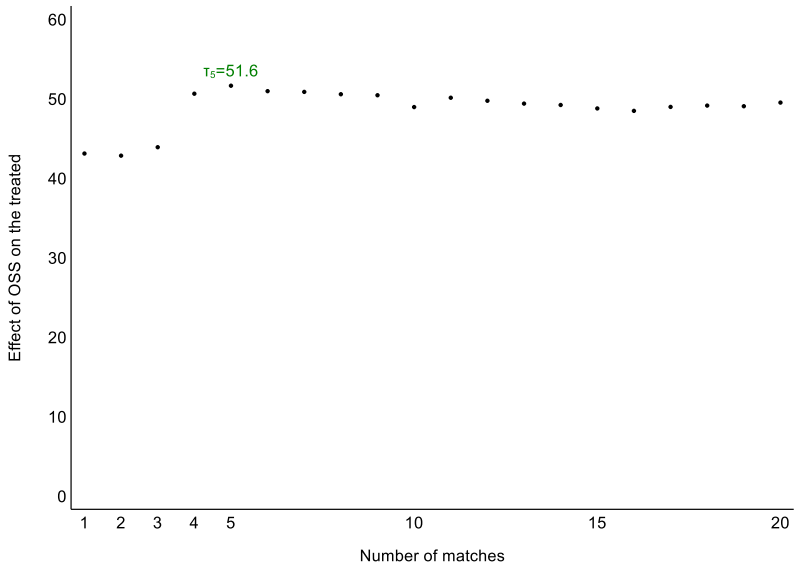


Figure 3 shows the distribution of the effect of OSS on speed for the treated ($\tau_{M,\text{treated}}^m$) from the bias-corrected matching estimator, with $M = 1, \ldots 20$. The highest value is with $M = 5$. A lower number of matches decreases the effect whereas a higher number does not change it much.

6. Policy recommendation and conclusion

The Covid-19 pandemic gives an opportunity to ask whether unrestricted access to publication and research data, material, reproducibility of results and open peer review could help scientists under pressure to work faster on solutions to find treatments. To our knowledge, no study so far examined the role of open-source science (OSS) on invention speed during periods as critical as the Covid-19 outbreak.

The present study asks whether researchers citing OSS invent faster than those not doing so. We employ a unique observational data set on projects submitted to *ClinicalTrials* during the early months of Covid-19 spread. Our results suggest that using OSS may accelerate invention by 51 days on average.

Although OSS was discussed in previous studies as a source for innovative process (Lariviere and Sugimoto, 2018; Woelfle et al., 2011), this study is the first to test the role of OSS. According to the report by OECD (2018), the development of a new drug generally takes an average of 10-15 years, which many projects fail along the path. Therefore, we plan to test a role of OSS on the success of the project of drug development in the pharmaceutical industry.

One cannot help but ask the question: why didn't academics emphasize more the potential impact of OSS? A possible argument relies on a dilemma of balancing between open science and intellectual proprietary (David, 2008). That is why recently governments collectively called for opening the access to research material, text, and dataset (PubMed, Kopernio, Dimensions). In addition, publishers encourage authors to publish open access material. See, e.g., Elsevier's Open Access Agreements with many research institutions in different countries. However, this strategy which mainly depends on sole action from governments is not sufficient, for collective actions are needed from individual, industry and academia. From a knowledge-sharing perspective (called 'outward science' in our paper), one can appeal to human nature as suggested by Woelfle et al. (2011). For instance, in the PZQ case study describes in that paper, the main motivation for private companies to provide free solutions simply was philanthropic nature of the project. Accordingly, it would be interesting to examine the differential impact of outward science on invention speed (relative to OSS). A further reason may be that funds substitute for OSS, a point that calls for further investigation.

From the perspective of fund receivers, a study of Lariviere and Sugimoto (2018) reveals that the rate and degree of open access for those projects receiving funds vary greatly across research domains (Engineering and technology, Health, Psychology, etc.) and funders (NIH, European Research Council, Economic and Social Research Council, and so on). Another research avenue is to take account of those dimensions when estimating the effect of OSS on speed of invention.

Finally, the interesting question for scientists living in less developed countries, searching for new knowledge, is how the benefits of OSS differ among those countries.

References

Abadie, A., Drukker, D., Herr, J., Imbens, G. (2004). Implementing matching estimators for average treatment effects in Stata. *The Stata Journal, 4*(3), 290-311.

Abadie, A., Imbens, G. (2011). Bias-corrected matching estimators for average treatment effects. *Journal of the American Statistical Association, 29*(1), 1-11.

Arthur, W. (2007). The structure of invention. *Research Policy*, *36*(2), 274-287.

Benjamin, D., Berger, J. et al. (2018). Redefine statistical significance. Nature human behaviour, *2*(1), 6-10.

Bercovitz, J., Feldman, M. (2011). The mechanisms of collaboration in inventive teams: Composition, social networks, and geography. *Research Policy*, *40*(1), 81-93.

Chesbrough, H. (2020). To recover faster from Covid-19, open up: Managerial implications from an open innovation perspective. *Industrial Marketing Management*, 88, 410-413.

Dasgupta, P., David, P. (1994). Toward a new economics of science. Research Policy, 23(5), 487-521.

David, P. (2008). The Historical Origins of 'Open Science': an essay on patronage, reputation and common agency contracting in the scientific revolution. *Capitalism and society*, *3*(2).

Dehejia, R., Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *The Review of Economics and Statistics, 84*(1), 151-161.

Fagerberg, J., Mowery, C., Nelson, R. (2004). *Innovation: A Guide to the Literature*. The Oxford Handbook of Innovation. Oxford: Oxford University Press.

Feldman, M., Dibaggio, L., Tran, T. (2023). The temporal value of local scientific expertise. *Industrial and Corporate Change, 32*(3), 647-672.

Galasso, A., Schankerman, M. (2015). Patents and cumulative innovation: Causal evidence from the courts. *The Quarterly Journal of Economics*, *130*(1), 317-369.

Gold, E. (2021). The fall of the innovation empire and its possible rise through open science. *Research Policy*, *50*(5), 104226.

Heirman, A., Clarysse, B. (2007). Which tangible and intangible assets matter for innovation speed in start-ups? *Journal of Product Innovation Management*, *24*(4), 303-315.

Hotez, P., Engels, D., Fenwick, A., Savioli, L. (2010). Africa is desperate for praziquantel. *The Lancet*, *376*(9740), 496-498.

Imbens, G. (2015). Matching methods in practice - Three examples. *The Journal of Human Resources, 50*(2), 373-419.

Imbens, G., Rubin, D. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge, USA: Cambridge University Press.

Khanagha, S., Volberda, H., Alexiou, A., Carmela Annosi M. (2022). Mitigating the dark side of agile teams: peer pressure, leaders' control, and the innovative output of agile teams. *Journal of Product Innovation Management*, *39*(3), 334-350.

Kessler, E., Chakrabarti, A. (1996). Innovation speed: A conceptual model of context, antecedents, and outcomes. *Academy of management review*, *21*(4), 1143-1191.

Larivière, V., Sugimoto, C. (2018). Do authors comply with mandates for open access? *Nature*, *562*(7728), 483-486.

Magalhães, J., Hartz, Z., Antunes, A., Maria do Rosário, O. (2017). An overview of the open science in times of big data and innovation to global health. *International Journal of Innovation*, *5*(3), 270-288.

Milan, E., Ulrich, F., Faria, L., Li-Ying, J. (2020). Exploring the impact of organisational, technological and relational contingencies on innovation speed in the light of open innovation. *Industry and innovation*, *27*(7), 804-836.

Munos, B., Chin, W. (2011). How to revive breakthrough innovation in the pharmaceutical industry. *Science translational medicine*, *3*(89), 89cm16.

OECD (2018). Pharmaceutical Innovation and Access to Medicines. OECD Health Policy Studies, OECD Publishing: Paris.

Noble, C., Spanjol, J. (2020). How are we faring? Reflections on coronavirus and its effects on the innovation management scholarly community. *Journal of Product Innovation Management*, 37(6), 474-482.

Rosenbaum, P., Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41-55.

Rosenbaum, P., Rubin, D. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, *39*(1), 33-38.

Rubin, D. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics, 2*(1), 1-26.

Stalk Jr, G., Hout, T. (1990). Competing against time. *Research-Technology Management*, *33*(2), 19-24.

Woelfle, M., Olliaro, P., Todd, M. (2011). Open science is a research accelerator. *Nature chemistry*, *3*(10), 745-748.

ABOUT OFCE

The Paris-based Observatoire français des conjonctures économiques (OFCE), or French Economic Observatory is an independent and publicly-funded centre whose activities focus on economic research, forecasting and the evaluation of public policy.

Its 1981 founding charter established it as part of the French Fondation nationale des sciences politiques (Sciences Po), and gave it the mission is to "ensure that the fruits of scientific rigour and academic independence serve the public debate about the economy". The OFCE fulfils this mission by conducting theoretical and empirical studies, taking part in international scientific networks, and assuring a regular presence in the media through close cooperation with the French and European public authorities. The work of the OFCE covers most fields of economic analysis, from macroeconomics, growth, social welfare programmes, taxation and employment policy to sustainable development, competition, innovation and regulatory affairs.

ABOUT SCIENCES PO

Sciences Po is an institution of higher education and research in the humanities and social sciences. Its work in law, economics, history, political science and sociology is pursued through ten research units and several crosscutting programmes.
Its research community includes over two hundred twenty members and three hundred fifty PhD candidates. Recognized internationally, their work covers a wide range of topics including education, democracies, urban development, globalization and public health.
One of Sciences Po's key objectives is to make a significant contribution to methodological, epistemological and theoretical advances in the humanities and social sciences. Sciences Po's mission is also to share the results of its research with the international research community, students, and more broadly, society as a whole.

PARTNERSHIP