RIETI Discussion Paper Series 25-E-089

Do Corporate Scientists Contribute to Firm Innovation? Empirical analysis by using linked dataset of research papers and patents in Japanese firms

MOTOHASHI, Kazuyuki
RIETI

TSUKADA, Naotoshi
RIETI

IKEUCHI, Kenta
RIETI

The Research Institute of Economy, Trade and Industry
https://www.rieti.go.jp/en/

RIETI Discussion Paper Series 25-E-089
September 2025

Do corporate scientists contribute to firm innovation? Empirical analysis by using linked dataset of research papers and patents in Japanese firms[1]

Kazuyuki Motohashi (RIETI, The University of Tokyo)
Naotoshi Tsukada (RIETI, University of Niigata Prefecture)
Kenta, Ikeuchi (RIETI)

Abstract

Corporate scientists that are involved in scientific activities, often leading to research paper publications, are important for corporate innovation, since science-based innovation tends to be transformative, spanning the boundaries of existing R&D pipelines. Such scientists can also play a role as a bridge between academic researchers, injecting scientific knowledge from outside the firm. However, the publication of internal corporate scientific activities could benefit competitor firms, providing them with input towards their own transformative innovation. In this study, we analyze this trade-off using a linked dataset of research papers and patents (disambiguated by paper author and patent inventor information and patent citation in research papers) of Japanese firms. Specifically, we analyzed two aspects, (1) contribution of corporate scientist research papers to in-house innovation (patent) and (2) capacity of corporate scientists to absorb scientific findings from outside their firms to obtain high quality patents. Our findings indicate that corporate scientists contribute to both aspects of innovation in their firms.

Keyword: corporate scientist, researcher level dataset, paper to patent citation

JEL codes: I23, O31, O34

The RIETI Discussion Paper Series aims at widely disseminating research results in the form of professional papers, with the goal of stimulating lively discussion. The views expressed in the papers are solely those of the author(s), and neither represent those of the organization(s) to which the author(s) belong(s) nor the Research Institute of Economy, Trade and Industry.

[1] This study is conducted as a part of the Project "Research on Digital Innovation Models" undertaken at the Research Institute of Economy, Trade and Industry (RIETI). The draft of this paper was presented at the RIETI DP seminar for the paper. I would like to thank participants of the RIETI DP Seminar for their helpful comments. The authors also acknowledge financial support from MEXT/JSPS KAKENHI (Grant Number: 23K25540)

1. Introduction

As the increasing importance of scientific knowledge in innovation can be observed across industries, such as pharmaceutical development (Pisano, 2006) and AI (Hartmann and Henkel, 2020), corporate scientists play a pivotal role in driving innovation and maintaining competitive advantage of a firm. The marginal productivity of research (R), as compared to development (D), has been evaluated, and most of literature suggests the R component contributes more to firm level productivity growth, due to the fact that spillovers from R part is greater than those from D part (Griliches, 1986; Akcigit et al., 2017). In addition, it is found that science based innovation tends to be more novel (Veugelers and Wang, 2019) and radical (Motohashi et. al, 2024), suggesting the importance of corporate scientists in firm's R&D activities.

However, knowledge spillover goes to not only to the firm conducting research activities, but also to its competitors. Since the research outputs are typically published as a research paper, which exposes its contents to outside the firm, and allows them to use it freely. In contrast, some inventions with specific industrial use can be patented, or can be kept as a trade secret, to prevent them from free riding by competitors. Arora et. al (2021) shows that the degree of firm's investment in R over D depends on the balance its internal use and the rent stealing effects by rivals.

This study investigates the role of corporate scientists, those conducting research activities inside a firm, in the corporate innovation, by using linked datasets of research papers and patents. These two documents are linked both at the researcher level (based on disambiguated author/inventor data, Ikeuchi et. al, 2024) and patent to research paper citation information (Marx and Fuegi, 2020; Marx and Fuegi, 2021). A corporate scientist is identified as an author of research paper(s) with a corporate affiliate, and the characteristics of research papers and patents, involving such person as an author/inventor has been investigated. Specifically, as for the research papers, the propensity of citation by patents is analyzed to see the degree of internal use of science, is analyzed. In addition, the innovation contribution of corporate scientist is studies by looking at the backward and forward citations of her patents.

In the next section, we review some related literature of our work. Subsequently, the data section, presenting the source and the methodology of linkage of research paper and patent, is provided. Then, the empirical analysis, both on the research papers and patents by corporate scientists, is presented. Finally, this paper concludes with summary of our observations with discussion and some policy/management implications.

2. Related Literature

It is well known that research component out of total R&D has higher rate of return than development one (Mansfield, 1980; Griliches, 1986). However, there are broader spillover effects by the research component (Acigit et al, 2007), which may allow greater rent stealing effects by product market rivals (Arora et al, 2021). Therefore, hiring corporate scientists who are engaged in research activity at firm is a double edged sword, in a sense of driving firm's research activity with higher rate of return and of helping the firm's competitors with its greater spillover effect.

Though the lens of resource based view, a corporate scientist can be a valuable resource generating tacit and complex knowledge, which it is difficult for the rivals to imitate (Herrera, 2020). In addition, science based innovation is generally a novel and radical one (Veugelers and Wang, 2019; Motohashi et al., 2024), so that a corporate scientist can expand a scope of R&D and contribute to strengthening a competitive position of its employer.

At the same time, a corporate scientist can constitute a firm level scientific knowledge base. A successful innovation can be achieved by series of assessing know technology opportunities, and interpreting them to assimilate a firm's internal R&D process. It is found that a research activity contribute to the efficiency of subsequent development stage (Nelson, 1959), and a corporate scientist often leads the collaboration and partnership activities with external knowledge sources, such as universities (Furukawa and Goto, 2006; Luo et al., 2009).

Finally, the network with scientific community is also an essential asset for the firm's innovation process. Publication of research papers and participation at academic conferences builds up a scientist's reputation in the community. A star scientist with substantial numbers of qualified papers can be a hub of information exchange of up-to-date scientific discovery, and the reputation in academic community facilitates effective research collaboration, particularly with academic scientists in universities and public research institutes. It is found that such stat scientist at firm contributes to firm's level organization capability generating high quality inventions (Grigoriou and Rothaermel, 2014).

As for another side of double edged sword of hiring corporate scientist is potential rent stealing effect by a competitor in product market. Bloom et al. (2013) distinguish firm's competitors in technology market and product market, to find that increase of R&D by former category firms benefits the rent of a focal firm, while increase of R&D by latter firms hurts it. Arora et al. (2021) further investigates this rent stealing effect by using NPL citation information, direct measurement of technology spillover of scientific papers, and finds some empirical evidences of rent stealing effects, that is, a firm invest more research (as compared to development) when it use research outputs (papers) more internally, and vice versa.

At researcher level, a firm needs to pay attention to leave of its corporate scientist. A star scientist with strong academic network is supposed to have substantial external option of its employer, particularly in academic field. Agarwal and Ohyama (2013) researches the career choice of researcher, including industry or academic and finds that a researcher choosing academia tends to be less sensitive in financial incentives, and more preference of independency. A corporate scientist tends to have an opposite personal preference at the timing of job selection, but such tastes may change over time. Therefore, an employer needs to know such characteristics in order to hire well qualified scientists and retain them. In the AI field, a firm has relatively generous policy for publication of research papers at academic conference, in order to hire large number of data scientists (Hartman and Henkel, 2020) . In addition, some firms have sponsored academic conferences in order to increase its visibility to academic community, as well as to make efficient hiring activities of corporate scientists (Baruffaldi and Poege, 2025).

3. Data

3-1. Data Source

In this paper, we use bibliometric information of research papers by OpenAlex, particularly the snapshot data on October 2022 (RELEASE 2022-10-10) (Priem et al., 2022). We have selected the papers in the journals in the list of Web of Science, SCIE (Science Citation Index Expanded), and the papers with at least one author whose address is in Japan.

For the patent information, we use IIP patent database (Goto and Motohashi, 2007), bases on JPO (Japan Patent Office) patent gazette, where at least one inventor's address in in Japan, as well (8,144,678 patents). The inventor affiliate information is obtained by using the applicant information within each patent, i.e., comparing the address of two records (if the inventor and applicant address is the same, the inventor's affiliate is the applicant of this patent) and use the applicant information as an inventor's affiliate for single applicant patents (Ikeuchi et al, 2024).

In this study, we use the document published after 1990 for both types (1,742,227 papers and 8,144,678 patents). In order for us to disambiguate institution names (universities, public institutions and firms), we used the NISTEP Dictionary of Names of Companies (NISTEP, 2022a) and the NISTEP Dictionary of Names of Universities and Public Research Institutions[1].

[1] The details of institution name harmonization of patent inventor/paper author affiliate are provided in Ikeuchi et al. (2024).

3-2. Author-inventor disambiguation dataset

In this study, we performed a large-scale disambiguation of inventors (patents) and authors (papers) to construct a unified person-level dataset. Records were first blocked by name, and then a supervised classifier was applied to estimate the probability that two records belong to the same person. The feature set ("Similarity Vector") was defined separately for patent–patent, paper–paper, and patent–paper pairs as follows:

A) Patent–Patent pairs

- Co-inventors' names: Jaccard coefficient.
- IPC class (3-digit): Jaccard coefficient.
- IPC subclass (4-digit): Jaccard coefficient.
- Applicants (NID/CID): Jaccard coefficient.
- Inventor affiliation (applicant of the inventor): same→1; either unknown→0.5; different→0.
- Inventor address proximity: 1 − (distance / 3,000 km).
- Abstract similarity: cosine similarity of 300-dimensional embeddings.

A) Paper–Paper pairs

- Co-authors' names: Jaccard coefficient.
- Affiliated institutions (NID): Jaccard coefficient.
- Address proximity: 1 − (distance / 3,000 km).
- Abstract similarity: cosine similarity of 300-dimensional embeddings.

B) Patent–Paper pairs

- Patent co-inventors vs. paper co-authors: Jaccard coefficient.
- Patent applicants vs. authors' affiliations (NID): Jaccard coefficient.
- Inventor's applicant (affiliation) vs. author's institution (NID): Jaccard coefficient.
- Address proximity (inventor vs. institution): max of 1 − (distance / 3,000 km).
- Abstract similarity: cosine similarity of 300-dimensional embeddings.

We evaluate GNB, LDA, QDA, Random Forest, AdaBoost, Gradient Boosting, and XGBoost, tuning via 5-fold CV with randomized search. By ROC-AUC, XGBoost performs best for all three pair types. Using predicted same-person probabilities as edges, we compared Connected Components, Label Propagation (NetworkX; LP-Net), Greedy Modularity Maximization, and DBSCAN (with probability-to-distance transform and precomputed distance). LP-Net attains the highest F1 (99.21%).

Within-name blocking is applied for efficiency; for paper records with first-name initials only, we link to the highest-probability cluster when the posterior ≥95%. Against independent KAKEN teacher data, F1 is 96.87% (paper–paper), 98.71% (patent–patent), and 81.64% (patent–paper). The cross-domain case mainly suffers from splitting due to English-name variants, as documented in the error analysis.

Table 1 reports the final results of our disambiguation. From a total of 25,477,918 patent and paper records, we identified 3,229,025 unique persons as inventors and/or authors. This comprehensive dataset enables us to consistently recognize individuals who engage in both scientific publications and patenting activities, providing a robust foundation for analyzing the role of corporate scientists in innovation.

	Total	Full Name info		
		Total records	Multiple records/name	One record/name
# of records (patent/paper * author/inventor)	25,477,918	25,444,389	24,989,363	455,026
Unique # of names (blocks)	1,425,143	1,425,143	970,117	455,026
Disambiguated Names	3,229,025	3,229,025	2,773,999	455,026
# of disambiguated names/# of unique names (per blocs)	2.266	2.266	2.859	1.000
Average number of documents per person	7.890	7.880	9.008	1.000

Table 1: Disambiguation Results

3-3. Patent to research paper citation

The linkage of paper and paper information can be measured by NPL (non patent literature, including research papers) citation by patent. In this paper, we use the Reliance on Science Dataset (https://relianceonscience.org/), matching the text record of NPL citations by the patent publication records of PATSTAT with research paper bibliometric information in the Open Alex (Marx and Fuegi, 2020; Marx and Fuegi, 2021).

The link with our dataset is straightforward for the research paper, since both of the datasets are developed by using the Open Alex with identical Open Alex paper id. In contrast, the linkage with patent by using JPO patent id gives only 37,799 (paper-patent) pairs, out of 47,844,586 pairs in total, since JPO examiners rarely use research paper as a citation document. Therefore, we use patent family information (docdb patent family) in PATSTAT to obtain equivalent patent

documents in the original paper-patent pairs. As a result, we have identified 640,538 paper-patent pairs in total, and 159,096 pairs with the papers authored by researchers in Japan.

4. Empirical Analysis

4-1. Descriptive statistics

Our author inventor disambiguated dataset enables us to analyze patenting activities (inventions with potential commercial application) of a corporate scientist (an author of research paper with corporate affiliate). In our dataset, there are 1.692,835 people who have at least one paper or patent with corporate affiliate. Among them, only 91,866 (5.4% of total) have at least one paper. Presumably, those people with any research paper have conducted some research activities at firm, so we call these people as a corporate scientist. It is found that a corporate scientist is minority among corporate staffs being involved with R&D activities. In addition, the majority of corporate scientists (54,153 out of 81,866, 58.9%) do not have any patent, so the division of labor between R and D is found within a firm.

		# of papers							
		0	1	2	3	4	5-9	10-	Total
# of patents	0	-	33,008	8,390	3,986	2,424	4,355	1,990	54,153
	1	640,281	2,097	688	331	240	522	287	644,446
	2	227,089	1,434	443	259	147	315	181	229,868
	3	132,068	1,022	371	177	107	230	123	134,098
	4	90,633	909	262	167	96	174	95	92,336
	5-9	224,588	3,101	1,074	556	311	651	375	230,656
	10-	286,310	9,105	3,728	2,121	1,378	2,663	1,973	307,278
	Total	1,600,969	50,676	14,956	7,597	4,703	8,910	5,024	1,692,835

Table 2 : Number of corporate author/inventors by paper and patent counts

4-2. Research papers by corporate scientists

In this section, the contribution of corporate scientists to firm's innovation is evaluated by the degree of their papers' citations to patents. Here, we use the samples of researchers with at least one paper with corporate affiliates (91.866 researchers). Out of 106,851 papers by those authors, only 5,723 papers (5.35%) have at least one patent citation. There can be multiple authors

per paper, these numbers are inflated for paper-author pairs, to 18,606 out of 253,422 pairs (7.34%).

Here, the regression analysis is conducted at this paper-author pair level. Since the corporate paper with patent citation is a minority group among all papers, we take two step approach. First, whether a paper has a patent citation or not is used as a dependent variable, then the number of patent citations are used only for the paper with at least one patent citation. We use the following four types of dependent variables.

- Probability (or numbers) of over all patent citation
- Probability (or numbers) of patent citation(s) within the same firm (self citation)
- Probability (or numbers) of patent citation(s) within the same firm, excluding the patents by the author of the paper (self citation, but not by yourself)
- Probability (or numbers) of patent citation(s), excluding self citation(s) (citation to others)

And the independent variables are as follows,

- Numbers of papers published for corporate scientist at the timing of paper publication, log scale (lpaper)
- Numbers of years after corporate scientist's first publication at firm, log scale (lage)
- Dummy variable whether corporate scientist has university collaboration paper at the timing of paper publication (UI_experience)

And the interaction terms of these variables. In addition, both paper's publication year and scientific field (20 concept classification of Open Alex) are controlled. Probit regression results of probability of each type of patent citation are presented in Table 3 and those of log patent citation counts for the papers with at least one patent citation are presented in Table 4.

(Table 3), (Table 4)

First, the more a corporate scientist publishes papers, the more likely her paper is cited by patents, reflecting her scientific activity's contribution to innovation. This finding is consistent to self citation probability and counts, suggesting more internal use of her scientific finding. However, the association between publication and patent citation is fading by a corporate scientist tenure at firm (negative and statistically significant coefficients to lage). Therefore, the focus of papers by a corporate scientist becomes far from commercialization as her scientific activities at firm goes by.

Second, a paper by the corporate scientist who has university collaboration experience by joint publication is less likely to be cited by patents. Again, a corporate scientist with joint a

research activity with university is more academic oriented, so that her output is less relevant to commercialization activity. This tendency increases by the number of past publications but decrease by the tenure at firm.

Table 5 shows the results of the fourth type of patent citation, citation to others. A corporate scientist activity is helping patenting of the parties other than her affiliated firm as well. The pattern of association with the number of past publication, the tenure and joint research experience with university is almost same as those of self citations.

(Table 5)

In order to investigate whether the corporate scientist's papers help to her own firm or the others, we use a dummy variable of internal use (self citation) or not as a dependent variable to regress by the same independent variables as before. The unit of observations of this analysis is patent-paper citation pair * numbers of authors of paper * numbers of inventors of patent * numbers of applicants of patent. The controlling variables here are the number of patent applications of affiliated firm and 4 types of dummy variables, paper publication year, paper field, affiliated firm's industry and patent application year.

As is shown in Table 6, as the tenure at firm increases, the paper less likely to be cited by her affiliated firm, but for those who have greater past publication, the age effect is reversed. Therefore, a corporate scientist, actively engaged with scientific activities at firm, contribute more to firm's patent even she stays at firm longer, and vice versa.

(Table 6)

In order to see the impact of a corporate scientist's past experience of joint research with university, we separate the number of past papers into two parts, the number of past papers with only corporate authors (log scale, lpaper_firm), and those with jointly authored with university (log scale, lpaper_univ). The association pattern of lpaper_firm is almost same as that of lpaper. In contrast, there is a negative association of lpaper_univ with self citation probability, which is not affected by a corporate scientist tenure.

4-2. Patents by corporation scientists

In this section, we analyze corporate scientist's capability to absorb scientific findings inside/outside firms for high quality patents. We suppose that participation of corporate scientists into the firm's R&D activities promotes the use of scientific knowledge and brings about high quality patents.

We construct samples by using three databases, our author-inventor disambiguated dataset, the PATSTAT database (2024 autumn), and the Reliance on Science Dataset (Marx and Fuegi). We focus on the patents of which application years are from 1990 to 2018, those of DOCDB patent families including at least one US patent application. The unit of observation is patent by inventor. In terms of a patent, there are multiple observations for the inventors. The observations are limited to inventors to whom the inventors' disambiguated IDs (CID) are assigned, and those of affiliate firms with organization IDs (NID). The number of observations is 1,426,753 (number of patents is 749,119, number of inventors is 369,141, average number of inventors per patent is 1.9). The dataset includes not only patents with NPL citations but also patents not citing NPL. The NPL citations information is obtained from the Reliance on Science Dataset. We suppose that patents not observed in the reliance on Science Dataset do not have any NPL citation.

In the next subsections, at first, we perform regression analysis using number of non-patent citations as the dependent variable to investigate whether participation of corporate scientists in research teams promotes the use of scientific knowledge in corporate R&D. We also do regression analysis using number of forward citations as the dependent variable to examine whether the use of science increases the value of patents. We use cumulative number of papers written by the inventor in firm at the timing of patent application to indicate a corporate scientist.

Variables	Explanation
nplall dnplall	No. of all NPL citations. Dummy which takes 1 if nplall >= 1, otherwise 0.
nplown dnplown	No. of NPL citations (citing patent NID and cited paper NID are same). Dummy which takes 1 if nplown >= 1, otherwise 0.
nplother dnplother	No. of NPL citations (citing patent NID and cited paper NID are different). Dummy which takes 1 if nplother >= 1, otherwise 0.
fc5	No. of forward citations (5 years window, docdb family count)

cpaper	Cumulative number of papers written by the inventor in firm at the timing of patent application.
cpat	Cumulative number of patents invented by the inventor in firm at the timing of patent application.
cpaper_other	Cumulative number of papers written by the inventor in other type organizations at the timing of patent application.
cpat_other	Cumulative number of patents invented by the inventor in other type organizations at the timing of patent application.
age_paper	No. of years since the inventor began publishing papers
age_pat	No. of years since the inventor began inventing patents
nb_inventors	No. of inventors of the patent
UI	Dummy takes 1 if the patent is invented through university-industry collaboration, otherwise 0.

Regarding the number of forward citations, in this study, the total number was counted, and no distinction was made between inventor/examiner citations, nor between self-citations and citations by other companies.

The basic statistics of the variables are in Table E. In addition to these variables, we control both patent application years and technical fields (WIPO 35 classifications).

(Table 7)

Figure 1 shows that the more papers a corporate scientist publishes, the more non-patent literature documents are cited in patents. It is also clear that there is a large difference in NPL citations between inventors with no papers and inventors with one or more papers.

Figure 1 Publications of corporate scientists and Number of NPL

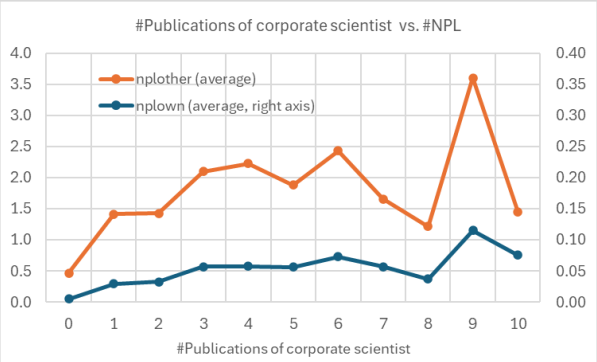
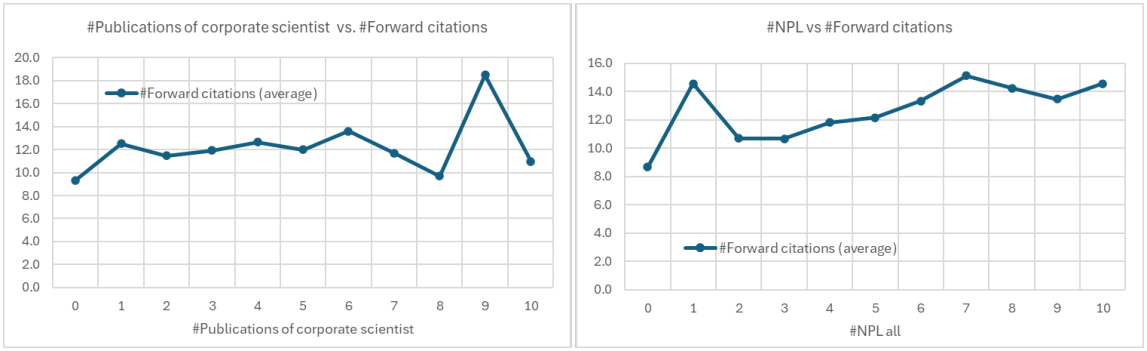


Figure 2 shows that patents by inventors with no published papers and inventors who do not cite NPL are likely to have fewer forward citations.

Figure 2. Use of science and Number of forward citations



4-3-1. Corporate scientists and their utilization of scientific knowledge

First, the more corporate scientists publish research papers, the more likely they are to cite scientific papers both of own firm and other organizations in their patents. Corporate scientists with longer scientific research periods at firms tend to cite more research papers. Therefore, this suggests that corporate scientists promote the use of internal/external scientific knowledge in R&D activities by conducting scientific research within their companies. In particular, it is found that the use of external scientific knowledge is driven by corporate scientists with more papers.

However, the promotional effects of corporate scientists are not always positive. The interaction effect between the number of their research papers and their experience period is statistically significantly negative. And as their number of patent inventions increases, utilization of scientific knowledge tends to diminish, likely due to the shift from basic research to more applied inventions.

Patents with more researchers involved in research, as well as patents invented through collaborations with universities, tend to utilize scientific knowledge more extensively.

(Table 8)

4-3-2. Use of scientific knowledge and patent quality

Next, we analyze whether the quality of patents, as measured by the number of forward citations, improves when corporate scientists participate in the firms' R&D

activities and use scientific knowledge. The estimation results of the Poisson regression is presented Table G. In equation (1) NPL are used as dummy variable, in equation (2) it used as count variable. In equations (3) and (4), NPL is divided into two categories: in-house papers and external papers.

In equation (1) the number of papers published by corporate scientists and the dummy variable of NPL significantly increase the number of forward citations. The interaction term is also positive, indicating that in R&D activities utilizing science knowledge, when corporate scientists with more paper publications participate in the research, patent quality measured by the forward citations is inclined to be high. However, taking into accounts that in equation (2) the interaction term is statistically significant and negative, so even in that case, an increase in NPL citations does not necessarily lead to an enhancement of forward citations.

Both the internal NPL dummy and the external NPL dummy are significantly positive when used as explanatory variables individually in equation (3). However, the interaction term between the number of paper publications by corporate scientists and the internal NPL dummy is significantly negative, indicating a negative impact on the number of forward citations. On the other hand, the interaction term with the external NPL dummy has a significantly positive coefficient. This suggests that corporate scientists with many academic papers may be more effective at absorbing external knowledge and contributing to corporate innovation.

The number of inventor's past patents and the number of inventors of the focal patent were estimated to be significantly positive, while the dummy variable of university-industry collaboration was estimated to be significantly negative, in equation (1)-(4).

(Table 9)

5. Conclusion

In this paper, a role of corporate scientist, a corporate employee with research paper publication, in its innovation, is investigated by looking at her research papers and patents. While only a minority of corporate papers are cited by subsequent patent, a paper by a corporate scientist with substantial numbers of past publications is likely to be cited by patents. However, the impact of past publications gets smaller, as her tenure at firm becomes longer. In addition, a paper by a researcher who have an experience of joint research with university is less likely to

be cited by patents. It is also found that a corporate scientist's paper contributes to subsequent patents not only by her own firm, but also by other organizations. The share of internal use of paper for patent become smaller, as the author's tenure gets longer, but this negative association is reversed when she has substantial numbers of past papers at her firm.

The patent analysis reveals that a corporate scientist is likely to cite research papers, not only by her own firm's but also by others. In addition, such patent with NPL citation are likely to be high quality one (greater numbers of forward citations), so that the absorptive capacity of corporate scientist, acquiring scientific knowledge from outside, could contribute to firm's level innovation activities. It is also found that the research publication record is positively correlated with NPL citations while such relationship is negative for the patent publication record. Moreover, the contribution of a corporate scientist's high quality patent can be found more for the one with substantial numbers of NPL citations from outside, instead of citations to the papers within her firm.

The foregoing analysis suggests the significance of a corporate scientist's role in firm's innovation in general, but also reveals the state of "double edged sword", i.e., research publications at firm helps, not her own firm, but also other parties, including her firm's rivals. An employer of such scientist has to allow her activities outside firm, such as publication and participation in academic conference, in order to strengthen the external network and her academic reputation to exploit vast opportunities in external scientific sources. But at the same time, the promotion of open activities could be backfired by its rivals' rent stealing effects.

This trade-off at firm is based on the non-rivalry nature of scientific knowledge. Due to the lack of effective instrument to appropriate the rent from the investment in science, some firm may be reluctant to do it. The recent trend of decline of corporate scientific investment could be explained by increasing market competition in high-tech industry (Arora et al., 2021). However, such behavior is not beneficial to the society over all. The substantial gap between the public and private rate of returns should be filled by policy support. For example, the R&D tax policy should be designed to make more incentives to R part, as compared to D part.

References

Ackigit, U., Hanley, D., and Serrano-Velarde, N. (2017), Back to basics, basic research spillovers, innovation policy and growth, NBER working paper 19473

Agarwal, R. and Ohyama, A., (2013), Industry or academia, basic or applied? Career choice and earnings trajectories of scientists, Management Science, 59(4): 950-970

Arora A., Belenzon S., and Sheer, L. (2021), Knowledge spillover and corporate investment in scientific research, American Economic Review, 111(3): 871-898

Baruffaldi, S. and Poege, F. (2025), Like Stars: How Firms Learn at Scientific Conferences, Management Science, 71(3): 1865-1888

Bloom, N., Shankerman, and van Reenen, J. (2013), Identifying technology spillovers overs and product market rivalry, Econometrics 81(4): 1347-93

Furukawa, R. and Goto, A. (2006), The role of corporate scientists in innovation, Research Policy, 35: 24-36

Goto, A., and Motohashi, K. (2007). "Construction of a Japanese Patent Database and a First Look at Japanese Patenting Activities," Research Policy, 36(9): 1431-1442.

Grigoriou, M. and Rothaermel, F. T. (2014), Structural micro-foundation of innovation, Journal of Management, 40: 586-615

Griliches, Z. (1986), Productivity, R&D and the basic research at firm level in the 1970's., American Economic Review 76(1): 141-54

Hartmann, P., & Henkel, J. (2020). "The rise of corporate science in AI: Data as a strategic resource," Academy of Management Discoveries, Vol. 6 No.3, : 359-381

Herrera, L. (2019), Effect of corporate scientists on firms' innovation activity : A literature survey, Journal of Economic Surveys, 34(1): 109-153

Luo, X. R., Koput, K. W., Powell, W. W. (2009), Intellectual capital or signal? The effect of scientists on alliance formation in knowledge intensive industries, Research Policy, 38: 1313-1325

Mansfield, E. (1980), Basic research and productivity increase in manufacturing, American Economic Review 70(5): 863-73

Marx, M. & Fuegi, A. (2020), Reliance on Science: Worldwide Front-Page Patent Citations to Scientific Articles" (2020), Strategic Management Journal 41(9):1572-1594

Marx, M. & Fuegi, A. (2022). "Reliance on Science by Inventors: Hybrid Extraction of in-Text Patent-to-Article Citations." Journal of Economics & Management Strategy 31 (2): 369–392.

Motohashi, K., Koshiba, H. and Ikeuchi, K. (2023). “Measuring Science and Innovation Linkage Using Text Mining of Research Papers and Patent Information,” RIETI Discussion Paper, 23-E-015.

Motohashi, K., Ikeuchi, K. and Yamaguchi, A. (2024). “Absorptive Capacity for Science-Based Innovation Propensity: An Empirical Analysis Using Japanese National Innovation Survey,” *The Journal of Technology Transfer*, forthcoming.

National Institute of Science and Technology Policy (2022a). Data on industrial research and development. The NISTEP Dictionary of Names of Companies ver2022_1, http://doi.org/10.15108/data_compdic001_2024_1

National Institute of Science and Technology Policy (2022b). Data on industrial research and development. The NISTEP Dictionary of Names of Universities and Public Research Institutes ver2022_1, http://doi.org/10.15108/data_rsorg001_2025_1

Nelson, R. R. (1959), The simple economics of basic science research, Journal of Political Economy, 67: 297-306

Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A Fully-open Index of Scholarly Works, Authors, Venues, Institutions, and Concepts. https://arxiv.org/abs/2205.01833

Veugelers, R. and Wang, J. (2019). “Scientific Novelty and Technological Impact,” *Research Policy*, 48(6) : 1362–137

Table 3 : Regression results of patent citation (dummy variable)

	Prob(citing patents)	Prob(citing patents)	Prob(citing patents)	Prob (Self citation)	Prob (Self citation)	Prob (Self citation)	Prob (Self-citation, not by yourself)	Prob (Self-citation, not by yourself)	Prob (Self-citation, not by yourself)
lpaper	0.560	0.581	0.637	0.514	0.533	0.578	0.463	0.481	0.533
	(41.23)**	(42.48)**	(37.67)**	(27.07)**	(27.92)**	(24.81)**	(22.72)**	(23.48)**	(21.46)**
lage	-0.060	-0.055	-0.104	-0.051	-0.045	-0.086	-0.030	-0.025	-0.072
	(6.58)**	(6.06)**	(8.28)**	(3.93)**	(3.50)**	(4.84)**	(2.15)*	(1.82)	(3.79)**
lage*lpaper	-0.072	-0.073	-0.070	-0.059	-0.060	-0.053	-0.054	-0.054	-0.046
	(13.17)**	(13.32)**	(11.80)**	(7.94)**	(8.03)**	(6.58)**	(6.63)**	(6.69)**	(5.23)**
UI experience		-0.140	-0.078		-0.141	-0.064		-0.134	-0.043
		(16.06)**	(4.01)**		(11.04)**	(2.25)*		(9.83)**	(1.42)
UI exp * lpaper			-0.095			-0.086			-0.100
			(5.29)**			(3.58)**			(3.90)**
UI exp*lage			0.072			0.049			0.055
			(5.49)**			(2.61)**			(2.79)**
_cons	-1.241	-1.247	-1.29	-2.165	-2.17	-2.209	-2.146	-2.151	-2.196
	(42.73)**	(42.87)**	(42.74)**	(45.57)**	(45.62)**	(45.21)**	(43.99)**	(44.03)**	(43.67)**
Pub year dummy	YES	YES	YES	YES	YES	YES	YES	YES	YES
Pub field dummy	YES	YES	YES	YES	YES	YES	YES	YES	YES
# of obs	214,881	214,881	214,881	214,865	214,865	214,865	214,878	214,878	214,878

Table 4 : Regression results of patent citation (numbers of forward citations)

	# of citing patents	# of citing patents	# of citing patents	Self citation counts	Self citation counts	Self citation counts	Self-citation, not by yourself	Self-citation, not by yourself	Self-citation, not by yourself
lpaper	0.146	0.149	0.168	0.048	0.049	0.042	0.027	0.030	0.016
	(21.13)**	(21.52)**	(20.50)**	(4.07)**	(4.15)**	(3.05)**	(1.73)	(1.88)	(0.88)
lage	-0.010	-0.009	-0.027	0.000	0.001	0.010	-0.003	-0.002	0.017
	(1.93)	(1.74)	(4.09)**	(0.01)	(0.06)	(0.87)	(0.22)	(0.16)	(1.17)
lage*lpaper	-0.028	-0.028	-0.027	-0.012	-0.012	-0.018	-0.008	-0.008	-0.017
	(9.84)**	(9.94)**	(8.66)**	(2.55)*	(2.57)*	(3.42)**	(1.24)	(1.27)	(2.49)*
UI experience		-0.023	-0.003		-0.008	-0.052		-0.017	-0.093
		(4.53)**	(0.310)		(0.910)	(2.71)**		(1.44)	(3.62)**
UI exp * lpaper			-0.033			0.026			0.049
			(3.81)**			(1.840)			(2.61)**
UI exp*lage			0.029			-0.001			-0.008
			(4.15)**			(0.070)			(0.500)
_cons	0.71	0.707	0.692	0.699	0.699	0.707	0.715	0.713	0.727
	(47.81)**	(47.61)**	(45.39)**	(21.67)**	(21.66)**	(21.62)**	(17.90)**	(17.83)**	(17.90)**
Pub year dummy	YES	YES	YES	YES	YES	YES	YES	YES	YES
Pub field dummy	YES	YES	YES	YES	YES	YES	YES	YES	YES
# of obs	21,030	21,030	21,030	6,185	6,185	6,185	4,249	4,249	4,249

Table 5 : Regression results of citations to others (dummy variable)

	Prob (citation to others)	Prob (citation to others)	Prob (citation to others)	Citation to others	Citation to others	Citation to others
lpaper	0.547	0.566	0.625	0.117	0.121	0.14
	(36.90)**	(37.94)**	(34.13)**	(15.42)**	(15.78)**	(15.66)**
lage	-0.05	-0.046	-0.098	-0.011	-0.01	-0.031
	(4.88)**	(4.49)**	(7.05)**	(1.96)	(1.80)	(4.10)**
lage*lpaper	-0.082	-0.083	-0.081	-0.021	-0.021	-0.019
	(13.44)**	(13.55)**	(12.35)**	(6.49)**	(6.57)**	(5.37)**
UI experience		-0.128	-0.075		-0.023	0.004
		(13.45)**	(3.52)**		(4.06)**	(0.34)
UI exp * lpaper			-0.097			-0.039
			(4.94)**			(3.98)**
UI exp*lage			0.084			0.030
			(5.77)**			(3.86)**
_cons	-1.343	-1.35	-1.393	0.725	0.722	0.706
	(43.91)**	(44.05)**	(43.78)**	(46.36)**	(46.16)**	(43.85)**
Pub year dummy	YES	YES	YES	YES	YES	YES
Pub field dummy	YES	YES	YES	YES	YES	YES
# of obs	214,881	214,881	214,881	15,756	15,756	15,756

Table 6 : Regression results of probability of self citations

	Self citation	Self citation	Self citation	Self citation	Self-citation, not by yourself	Self-citation, not by yourself	Self-citation, not by yourself	Self-citation, not by yourself
lpaper	0.033	-0.048			-0.025	-0.149		
	(1.92)	(1.66)			(1.28)	(4.47)**		
lage	-0.047	-0.096	-0.040	-0.083	-0.016	-0.091	-0.020	-0.085
	(3.19)**	(4.68)**	(2.86)**	(4.77)**	(0.97)	(3.93)**	(1.23)	(4.27)**
lage*lpaper		0.041				0.064		
		(3.43)**				(4.61)**		
lpaper_firm			0.025	-0.071			0.006	-0.133
			(1.75)	(2.68)**			(0.36)	(4.37)**
lpaper_univ			-0.002	-0.015			-0.054	-0.112
			(0.15)	-0.49			(3.53)**	(3.27)**
age*lpaper_firm				0.052				0.071
				(4.42)**				(5.33)**
age*lpaper_univ				-0.002				0.018
				(0.13)				(1.20)
lpcount	0.025	0.025	0.025	0.025	0.016	0.016	0.016	0.016
	(4.40)**	(4.34)**	(4.30)**	(4.32)**	(2.50)*	(2.44)*	(2.39)*	(2.37)*
_cons	-1.632	-1.544	-1.611	-1.514	-2.174	-2.059	-2.179	-2.05
	(12.61)**	(11.73)**	(12.53)**	(11.59)**	(13.00)**	(12.17)**	(13.09)**	(12.17)**
Publication dummy	31557	31557	31557	31557	28423	28423	28423	28423
Paper field dummy	YES	YES	YES	YES	YES	YES	YES	YES
Industry dummy	YES	YES	YES	YES	YES	YES	YES	YES
App year dummy	YES	YES	YES	YES	YES	YES	YES	YES
# of obs	29,020	29,020	29,020	29,020	29,020	26,327	26,327	26,327

Table 7: Basic statistics

Variable	Obs	Mean	Std. dev.	Min	Max
fc5	1,426,753	9.598	15.487	0	1,109
nplall	1,426,753	0.585	3.663	0	569
nplown	1,426,753	0.009	0.127	0	15
nplother	1,426,753	0.577	3.580	0	569
nb_inventors	1,426,753	3.465	2.330	1	38
UI	1,426,753	0.003	0.050	0	1
cpaper	1,426,753	0.336	2.377	0	188
cpat	1,426,753	45.479	268.037	1	15,127
cpaper_other	1,426,753	0.566	7.507	0	1149
cpat_other	1,426,753	4.881	15.318	0	612
age_paper	1,426,753	0.629	2.870	0	28
age_pat	1,426,753	7.858	6.737	0	28

Table 8: Regression results of NPL citations of patents

	(1)		(2)		(3)		(4)		(5)		(6)	
	Poisson		Poisson		Poisson		Probit		Probit		Probit	
VARIABLES	npl_all		npl_own		npl_other		Prob(npl_all>0)		Prob(npl_own>0)		Prob(npl_other>0)	
	(coef.)	(marg.)	(coef.)	(marg.)	(coef.)	(marg.)	(coef.)	(marg.)	(coef.)	(marg.)	(coef.)	(marg.)
ln_cpaper	1.007***	0.589***	1.587***	0.0137***	0.992***	0.572***	0.421***	0.0748***	0.748***	0.00560***	0.421***	0.0747***
	(0.00449)	(0.00271)	(0.0266)	(0.000261)	(0.00456)	(0.00270)	(0.0101)	(0.00178)	(0.0217)	(0.000179)	(0.0101)	(0.00178)
ln_cpat	0.392***	0.229***	0.694***	0.00600***	0.388***	0.224***	0.0243***	0.00431***	0.0260*	0.000194*	0.0250***	0.00442***
	(0.00243)	(0.00144)	(0.0218)	(0.000196)	(0.00244)	(0.00143)	(0.00331)	(0.000588)	(0.0140)	(0.000105)	(0.00332)	(0.000588)
ln_age_paper	0.0627***	0.0367***	0.0430**	0.000372**	0.0649***	0.0374***	0.00173	0.000308	0.00429	3.21e-05	0.00273	0.000483
	(0.00264)	(0.00155)	(0.0186)	(0.000161)	(0.00267)	(0.00154)	(0.00473)	(0.000839)	(0.0131)	(9.79e-05)	(0.00473)	(0.000839)
ln_age_pat	-0.369***	-0.216***	-0.210***	-0.00181***	-0.371***	-0.214***	-0.134***	-0.0239***	-0.00472	-3.53e-05	-0.133***	-0.0236***
	(0.00235)	(0.00140)	(0.0202)	(0.000175)	(0.00237)	(0.00139)	(0.00336)	(0.000597)	(0.0123)	(9.19e-05)	(0.00336)	(0.000596)
ln_cpaper * ln_age_paper	-0.308***	-0.180***	-0.404***	-0.00349***	-0.306***	-0.176***	-0.107***	-0.0190***	-0.159***	-0.00119***	-0.108***	-0.0192***
	(0.00170)	(0.00102)	(0.0104)	(9.49e-05)	(0.00173)	(0.00102)	(0.00388)	(0.000688)	(0.00862)	(6.65e-05)	(0.00388)	(0.000688)
ln_cpat * ln_age_pat	0.0108***	0.00632***	-0.0977***	-0.000844**	0.0121***	0.00695***	0.0281***	0.00498***	-0.0136***	-0.000102**	0.0277***	0.00490***
	(0.000864)	(0.000506)	(0.00725)	(6.31e-05)	(0.000871)	(0.000502)	(0.00129)	(0.000228)	(0.00502)	(3.76e-05)	(0.00129)	(0.000228)
ln_cpaper_other	0.105***	0.0615***	0.249***	0.00215***	0.102***	0.0587***	0.0313***	0.00556***	0.137***	0.00103***	0.0311***	0.00551***
	(0.00177)	(0.00104)	(0.0111)	(9.76e-05)	(0.00179)	(0.00103)	(0.00299)	(0.000532)	(0.00677)	(5.24e-05)	(0.00300)	(0.000531)
ln_cpat_other	-0.0991***	-0.0580***	-0.143***	-0.00123***	-0.0983***	-0.0567***	0.0594***	0.0105***	-0.00139	-1.04e-05	0.0590***	0.0105***
	(0.00106)	(0.000623)	(0.00911)	(7.95e-05)	(0.00107)	(0.000618)	(0.00127)	(0.000226)	(0.00511)	(3.83e-05)	(0.00127)	(0.000225)
ln_nb_inventors	0.120***	0.0701***	0.117***	0.00101***	0.120***	0.0690***	0.152***	0.0269***	0.123***	0.000917***	0.151***	0.0267***
	(0.00174)	(0.00102)	(0.0147)	(0.000128)	(0.00176)	(0.00102)	(0.00226)	(0.000402)	(0.00947)	(7.20e-05)	(0.00226)	(0.000401)
UI	0.640***	0.375***	1.673***	0.0145***	0.608***	0.351***	0.220***	0.0390***	1.099***	0.00822***	0.199***	0.0352***
	(0.0109)	(0.00641)	(0.0523)	(0.000470)	(0.0112)	(0.00647)	(0.0234)	(0.00415)	(0.0337)	(0.000273)	(0.0236)	(0.00418)
Constant	-2.209***		-9.637***		-2.201***		-1.393***		-3.910***		-1.394***	
	(0.0133)		(0.323)		(0.0133)		(0.0144)		(0.132)		(0.0144)	
App year dummies	Yes		Yes		Yes		Yes		Yes		Yes	
Tech field dummies	Yes		Yes		Yes		Yes		Yes		Yes	
Observations	1,426,148		1,426,148		1,426,148		1,426,148		1,405,283		1,426,148	

Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table 9: Regression results of forward citations

	(1)		(2)	
	Poisson		Poisson	
VARIABLES	fc5		fc5	
	(coef.)	(marg.)	(coef.)	(marg.)
ln_cpaper	0.0140***	0.134***	0.0550***	0.528***
	(0.000756)	(0.00725)	(0.000712)	(0.00683)
dnplall	0.422***	4.052***		
	(0.000789)	(0.00765)		
ln_cpaper * dnplall	0.0464***	0.445***		
	(0.00113)	(0.0108)		
ln_nplall			0.251***	2.410***
			(0.000494)	(0.00479)
ln_cpaper * ln_nplall			-0.0444***	-0.426***
			(0.000543)	(0.00521)
ln_cpat	0.124***	1.185***	0.123***	1.177***
	(0.000231)	(0.00224)	(0.000232)	(0.00225)
ln_nb_inventors	0.246***	2.362***	0.249***	2.393***
	(0.000420)	(0.00408)	(0.000421)	(0.00409)
UI	-0.394***	-3.779***	-0.481***	-4.620***
	(0.00630)	(0.0605)	(0.00631)	(0.0605)
Constant	1.292***		1.313***	
	(0.00344)		(0.00344)	
App year dummies	Yes		Yes	
Tech field dummies	Yes		Yes	
Observations	1,426,148	1,426,148	1,426,148	1,426,148

Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

	(3)		(4)	
	Poisson		Poisson	
VARIABLES	fc5		fc5	
	(coef.)	(marg.)	(coef.)	(marg.)
ln_cpaper	0.0141***	0.135***	0.0436***	0.418***
	(0.000762)	(0.00731)	(0.000750)	(0.00719)
dnplown	0.322***	3.093***		
	(0.00540)	(0.0518)		
dnplother	0.422***	4.054***		
	(0.000790)	(0.00766)		
ln_cpaper * dnplown	-0.0581***	-0.558***		
	(0.00408)	(0.0391)		
ln_cpaper * dnplother	0.0446***	0.428***		
	(0.00113)	(0.0108)		
ln_nplown			-0.302***	-2.899***
			(0.00385)	(0.0370)
ln_nplother			0.272***	2.606***
			(0.000561)	(0.00543)
ln_cpaper * ln_nplown			-0.0330***	-0.317***
			(0.00328)	(0.0314)
ln_cpaper * ln_nplother			-0.0206***	-0.198***
			(0.000816)	(0.00783)
ln_cpat	0.124***	1.187***	0.123***	1.181***
	(0.000231)	(0.00224)	(0.000232)	(0.00225)
ln_nb_inventors	0.246***	2.360***	0.248***	2.384***
	(0.000420)	(0.00408)	(0.000421)	(0.00409)
UI	-0.414***	-3.977***	-0.471***	-4.524***
	(0.00632)	(0.0606)	(0.00631)	(0.0605)
Constant	1.293***		1.307***	
	(0.00344)		(0.00344)	
App year dummies	Yes		Yes	
Tech field dummies	Yes		Yes	
Observations	1,426,148	1,426,148	1,426,148	1,426,148

Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1