RIETI Discussion Paper Series 15-E-105

The Use of Science for Inventions and its Identification: Patent level evidence matched with survey

NAGAOKA Sadao
RIETI

YAMAUCHI Isamu
RIETI

Research Institute of Economy, Trade & Industry, IAA

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

RIETI Discussion Paper Series 15-E-105
August 2015

The Use of Science for Inventions and its Identification: Patent level evidence matched with survey†

NAGAOKA Sadao *
Research Institute of Economy, Trade and Industry/ Tokyo Keizai University

YAMAUCHI Isamu **
Research Institute of Economy, Trade and Industry

Abstract

While backward citation information disclosed in patent documents is often used for tracing the scientific sources of innovations, it is still poorly understood how well the backward citations trace the actual knowledge flow from science. This paper directly evaluates both the completeness and the noise of the inventor citation information, linking the results of an original inventor survey on scientific sources to the dataset of non-patent literatures (NPLs) revealed in the entire patent document. We find that patent citations to NPLs are not only noisy but also highly incomplete. More important science sources are not necessarily more revealed. However, controlling for the propensity to cite NPLs, our estimation results show that the revealed NPLs are more likely to predict the existence of important scientific sources when the inventor refers to highly cited scientific literature early after its publication. We also find that the NPLs revealed at the place where an invention is described provide important additional information in identifying science sources.

Keywords: Citation, Patent, Scientific source, Non-patent literature
JEL classification numbers: O31, O32, O34

---

† This paper is partly based on the research project "Research on Innovation Process and its Institutional Infrastructure" undertaken at Research Institute of Economy, Trade and Industry (RIETI) and "Inventor Survey on the Scientific Sources of Inventions" commissioned by Japan Science and Technology Agency (JST). The survey "Inventor Survey on the Scientific Sources of Inventions" was conducted by RIETI. The authors would like to thank Naotoshi Tsukada, Yusuke Naito, Takahiro Maeda and Junko Morikawa for their assistance and contribution to construct dataset. The authors are grateful for helpful comments and suggestions by Masahisa Fujita, Masayuki Morikawa and Discussion Paper seminar participants at RIETI.

* Tokyo Keizai University, *Email address*: nagaoka@tku.ac.jp

** Corresponding author. 1-3-1 Kasumigaseki Chiyoda-ku Tokyo 100-8901, Tel.:+81 3 3501 1393, *Email address*: yamauchi-isamu@rieti.go.jp

1. Introduction

Measuring the knowledge flow from science to invention is critically important for assessing the effectiveness of science policy for promoting innovation. Backward citation information disclosed in patent documents is frequently used to trace the knowledge flow in the innovation research. However, in the patent document, the prior art is disclosed only to clarify the novelty and the inventive step of the invention, so that it does not necessarily indicate the scientific source. Patent citations disclosed as prior art is bound to be *incomplete*, since the inventor is not expected to disclose the scientific source when it is irrelevant to the novelty or the inventive step of the inventions. At the same time, patent citations disclosed as prior art may not inform the knowledge sources for the invention, either because prior art is added by an inventor only for explaining the novelty of the invention or because such prior art is added ex-post by examiners or patent attorney. As a result, patent citation is also bound to be a *noisy* indicator for identifying the actual knowledge flow of invention. For an example, discoveries related to drug development are often supported by scientific progress in identifying the target molecule and in the screening methods of drug candidates. On the other hand, the patentability requires that the newly discovered drug molecule should be new and significantly different from the existing molecules. In this case, the inventor discloses the existing related molecules as prior art but not necessarily refer to the targets nor to the screening method enabling his/her research due to scientific progress as prior art. Therefore, the prior art disclosure in this case is totally incomplete and contains only noises for understanding the scientific source.

While these problems may be well recognized, there are no systematic data and analysis at the project level. This paper examines the underlying mechanism of the revelation of scientific sources of inventions, which enables us to understand the accuracy, both the completeness and the noise, of the citation data to trace knowledge flow. We implemented an inventor survey to collect detailed information on the scientific sources for inventions. Matching the survey information to the non-patent literatures ("NPLs", hereafter) extracted from the entire part of the patent documents (covering not only the prior art part but also the part describing the invention), we examine the accuracy of the citation data and also assess how we may use the characteristics of cited NPLs to predict the existence of scientific sources. To the best of our knowledge, this paper is the first systematic attempt to directly evaluate the accuracy of the inventor citation information, based on a survey at an individual patent level, with respect to scientific sources.

Most of the related existing studies use only the citation information disclosed as prior art. However, this paper covers not only the prior art but also the literatures referred to in order to explain the invention (including the method of obtaining the product in the case of product patent and the utility of the invention), which accounts for more than 40 % of the NPLs cited in patent documents (see section 3). Through text mining we extracted all the literatures revealed in the patent document. Therefore, the coverage of citation information in this paper is beyond the requirement of patent law on prior art disclosure, and thus is more comprehensive than the previous studies. We differentiate “revelation” from “disclosure”: the latter is for meeting the requirement of patent law while the former is for identifying the knowledge sources.

Jaffe et al. (2000) assessed the relevancy of cited patent information with knowledge flow based on survey data. However, they do not identify the actual specific knowledge that contributed to the invention. Therefore, they do not evaluate the “completeness” of the citations, which is perhaps more important than “noise” of the citations in tracing the knowledge flow. Furthermore, their sample does not cover NPLs. Roach and Cohen (2013) analyze both “errors of omission” and “errors of commission” of citations as a measure of knowledge flows from public research. They find that patent citations reflect codified (public) knowledge flow but may fail to capture private and contract-based knowledge flow. They also point out the influence of patenting and citation strategies of a firm on citation flows. However, their analysis is based on the R&D lab level, and does not directly identify the match between the knowledge flow and the citations disclosed in the patent documents at a project level. Moreover, both of these two studies do not distinguish examiner citations from inventor citations and depend only on prior art citations.

This paper directly evaluates the accuracy of the inventor citation information, linking the results of an original inventor survey on scientific sources to the dataset of NPLs revealed in the entire patent document. Our survey results indicate that for about a quarter of inventions, scientific knowledge embodied in literature is *essential* to envisage or implement the R&D. Furthermore, comparing the scientific literatures indicated by respondents to the survey and the NPLs revealed in a patent, we find that only 17% of the inventions with important scientific sources reveal such important sources in their patent documents (37 % if we include the ambiguous cases). More important science sources are not necessarily more revealed. We also find that 82% of the inventions with citation to NPLs reveal only unimportant literatures (61% if we do not count the ambiguous cases). These results indicate that the patent citations are incomplete and noisy indicator in tracing the actual knowledge flow. However, our estimation results show that the citation information is still a useful index to trace the knowledge flow. We find that revealed NPLs as prior art are more likely to predict the existence of scientific literatures as important knowledge sources when the inventor refers to highly cited literature early after its publication. This result suggests that we can partially predict the existence of scientific sources by the inventor citation information. Moreover, our results show that the NPLs revealed at the place where an invention is described significantly add information on the knowledge sources (43% of the important scientific literatures are revealed in this part).

The rest of the paper is organized as follows. Section 2 reviews related studies. Section 3 describes the design of the inventor survey and explains the datasets constructed. Section 4 and 5 provide the empirical results. Section 6 concludes the paper.

2. Related studies

There are seminal studies that analyse the scientific sources of innovation, such as Mansfield (1995), and Klevorick et al. (1995). However, most of these studies are based on firm-level questionnaire surveys. Since knowledge sources for the inventions in a firm are likely to be highly heterogeneous and each individual inventor is unlikely to have good information on the knowledge source of the inventions by the others, such surveys may not be able to comprehensively identify the science sources. Given such heterogeneity, it is also difficult to evaluate the degree to which the disclosed prior art in patent documents reflects the knowledge flow, based on firm-level data. Therefore, we will use invention-level data on the scientific sources in our analysis.

Early literatures on patent citation data made clear that backward citation information disclosed in the front page of the U.S. patent document includes large noise (e.g. Jaffe et al., 1998). Through a questionnaire survey, Jaffe et al. (2000) find that one third of inventors did not recognize the literature cited in the patent document as knowledge sources of the invention before receiving the survey, whereas only 40% of inventors recognized it either before or during the development of their own invention. One fundamental reason for such noise is that the front page of the U.S. patent document aims at informing the public all relevant prior art to clarify the scope of the patent, irrespective of whether the inventor knows such art when his invention is made. In particular, a large proportion of such citations are added by an examiner, complementing the disclosure made by the applicant. According to Alcacer and Gittelman (2006) (see also Alcacer et al.; 2009), two thirds of the citations were added by the examiners, and the citations were made only by the examiners for approximately 40% of patents, for the US patents issued from 2001 to August 2003. Our study uses only the inventor citations disclosed in the Japanese patent application documents, in order to reduce the noises added ex-post by examiners.

Only a few existing studies quantitatively evaluate the incompleteness of the citations. Roach and Cohen (2013) compares the number of the backward citations from corporate inventions to public research outcomes with a survey measure on the proportion of firm's R&D projects using public research outcomes, at R&D laboratory level. They find that relying on the patent citations underestimates the contribution of the public research (patent citation is an incomplete measure of knowledge flow). In addition, they find that the firm's patenting and disclosure strategies affect the number of citations, which can decrease the accuracy of citation information. One constraint of their study is that their survey does not measure the intensity or the contribution of public research in the industrial R&D project but only the breadth of its use, given that their survey unit is a laboratory. Another is that they do not exclude examiner citations. Duguet and MacGarvie (2005) analyse whether the EPO citations (examiner citations) are a good measure of knowledge flows, based on the French innovation survey data. They find that backward and forward citations are related to firms' statements about their acquisition and dissemination of new technology respectively. Their focus is not knowledge flow from science (in fact, they find strong correlations between backward citations and equipment purchase). Moreover, their measure of knowledge flow is quite limited: whether a particular channel is used by a firm for acquiring new technology. More recently, Nelson (2009) provides an analysis based on the impact of a very specific scientific discovery: the recombinant DNA (rDNA) technology. He collected information on patent citations, licenses, and publication citations for the single invention and compares the coverage of these three indices over knowledge spill-over. He shows that direct patent citations miss a large proportion of licensing organizations which actually released related products. Moreover, publication citations are more effective in picking up universities/ PROs than patent citations. Thus, the patent citation is a very incomplete measure of knowledge flow in this case. While highly informative, this study is based on a single case.

Unlike existing studies, this paper assesses directly how patent backward citations trace knowledge flow from science at project level, rather than at laboratory or firm level, covering all technology fields. Through the inventor survey, we collected detailed information on the contribution of scientific sources to a specific invention as well as their identifications (such as the title of the literature, its author, and the date of publication). Correlating such specific information on the scientific sources to the NPLs revealed by the inventor in the patent document, we can examine whether the inventor refers to these important scientific literatures in patent documents. This procedure can rule out any spurious correlations which could emerge when the count data is used. As for citation data, we use only inventor citations, excluding all ex-post citations by examiners but including those made in the part describing the invention, in the patent application documents which are automatically disclosed in 18 months from priority date.

3. Data construction and overview

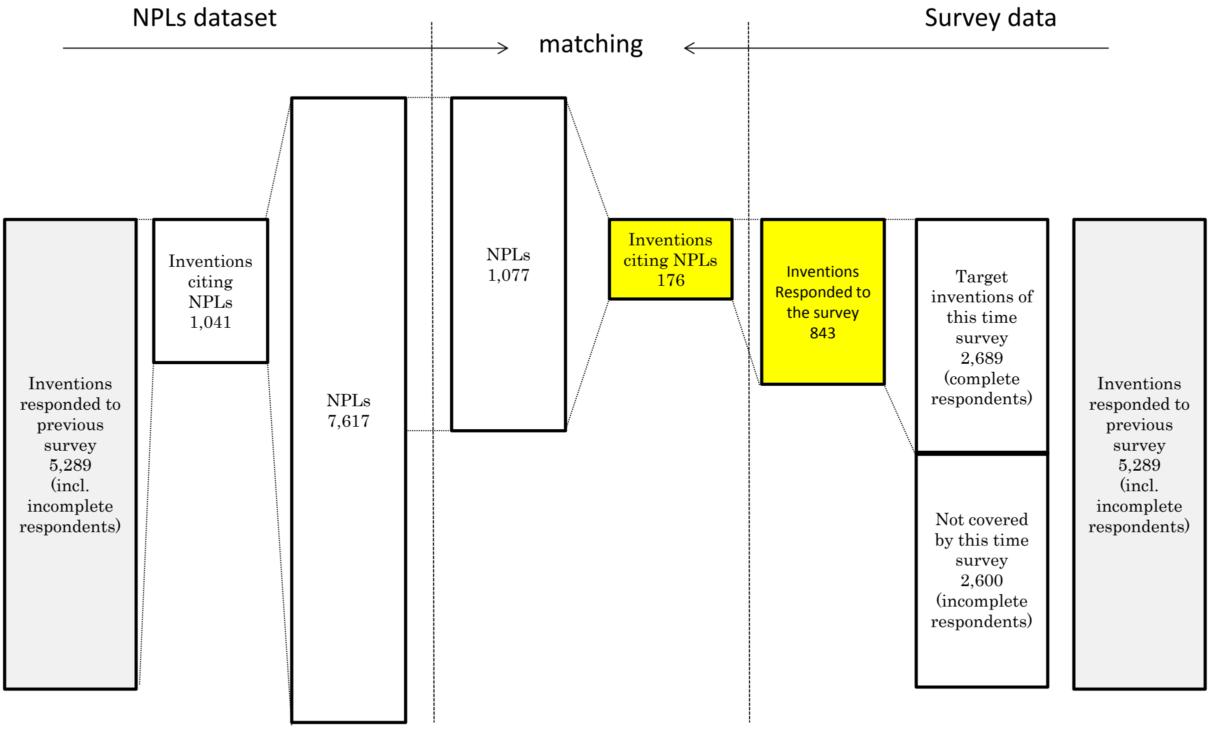
3.1 NPLs dataset and inventor survey

Our analysis is based on two datasets. The first comprises the survey responses, and the other comprises NPLs. Figure 1 illustrates the relation of these two data sources. The population of the survey is the responses to the previous survey conducted in 2010[1]. Before conducting this follow-up survey, we constructed a dataset comprising NPLs. First, we extracted all the NPLs revealed in the whole body of patent application documents of the population, 5,289 inventions, through text mining. In this process, we identified the places where the literatures are revealed: the place where prior art is disclosed, and the place where the invention is described. Consequently, 7617 NPLs were extracted. Thereafter, we merged the data on the Web of Science and the Japanese literature database provided by the Japan Science and Technology Agency (JST) to these 7,617 NPLs. We collected information on the author name, affiliation, journal name, publication date, and title of the paper. If there were unmatched data, we manually collected information via web search.

After the construction of the NPLs dataset, we conducted the questionnaire survey, which targeted the respondents of the previous inventor survey. The previous survey targeted 17,000 patent applications filed at both the European Patent Office (EPO) and Japan Patent Office (JPO) with a priority date of 2003-2005. It received 5,289 complete or incomplete responses. The 2,689 inventors who provided their email addresses in the previous survey were the eventual targets of this survey. The procedure used is a web survey and the number of responses was 843. This number is the final sample of this follow-up survey. Only 176 out of 843 inventions (20.9%) refer to at least one NPL in the whole body of patent document; these 176 inventions cite 1,077 NPLs in total.

[1] Detailed information on the previous inventor survey is given in Nagaoka et al. (2012).

Figure 1. Relation between the NPLs dataset and the inventor survey



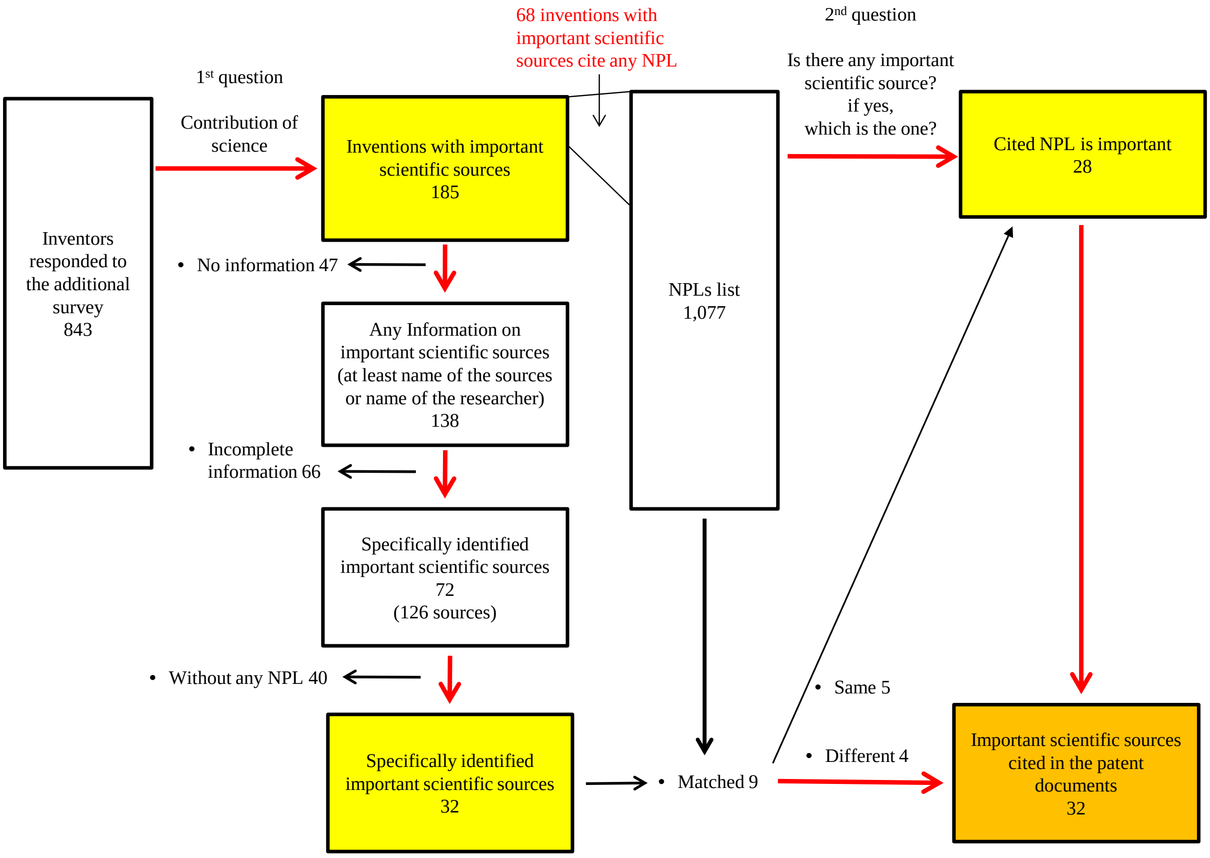
3.2 Survey strategy

Figure 2 illustrates the strategy of the survey. First, before showing the list of NPLs to the respondents, we asked whether there were any scientific source that were essential for the conception or implementation of R&D to ensure that the answers are not affected by the list of NPLs. Of 843 inventions, 185 have important scientific sources. Thereafter, we asked for the detailed information on the important scientific sources (e.g., the name of the researcher who developed the scientific source, his/her affiliation, the title of the paper). 138 respondents provided at least the name of the scientific sources, of which 72 gave us the detailed information on scientific sources. Of 72 inventions, we identified 32 inventions that we specified scientific literatures using the Web of Science and Japanese literature database.

After receiving the answer on scientific sources, we displayed the list of NPLs disclosed in the respondents' patent document and asked whether those literatures were important for conception or implementation of R&D. In total, only 28 inventions (out of 68 inventions with important scientific source) have important NPLs

Matching the specifically identified scientific sources to the important NPLs cited in the patent documents enabled us to identify 32 inventions revealing the important scientific sources in the patent documents.

Figure 2. Survey strategy



3.3 Data overview

Table 1 shows the distribution of our sample and the number of revealed NPLs by the technology field based on the ISI-OST-INPI classification. Moreover, we show the share of NPLs revealed at the places where prior art is disclosed and where invention is described, respectively. We find that the number of revealed NPLs per patent application is outstanding in the fields of Biotechnology, Organic Chemicals, and Pharmaceuticals. Furthermore, Table 1 indicates that approximately 55% of 1,077 NPLs are cited at the place where prior art is disclosed, and the rest (45%) are cited where the invention is described. Therefore we find that covering the NPLs revealed at the place where an invention is described provide significant additional information to trace the knowledge flow.

Table 1. Number of cited NPLs by technology field and the place in the patent document

	Number of inventions	Number of cited NPLs	Number of cited NPLs per patent application	Share of the NPLs cited at the place where prior art is described (%)	Share of the NPLs cited at the place where invention is described (%)
Biotechnology	15	222	14.8	33.8%	66.2%
OrganicChem	57	419	7.4	60.1%	39.9%
Pharmaceuticals/Cosmetics	29	134	4.6	57.5%	42.5%
Materials	23	28	1.2	71.4%	28.6%
Polymers	49	54	1.1	64.8%	35.2%
MedicalTechn	22	17	0.8	35.3%	64.7%
Semiconductors	38	23	0.6	26.1%	73.9%
SurfaceTechn	19	11	0.6	45.5%	54.5%
Analysis/Measurement/ControlTechn	71	40	0.6	90.0%	10.0%
Optical	31	16	0.5	68.8%	31.3%
Electr/Energy	67	32	0.5	43.8%	56.3%
Audiovisual	35	15	0.4	93.3%	6.7%
ChemEngineering	19	8	0.4	75.0%	25.0%
PetrolChem/materialsChem	17	7	0.4	71.4%	28.6%
Environment	14	5	0.4	60.0%	40.0%
MechElements	36	10	0.3	80.0%	20.0%
Matprocessing/Textiles/Paper	41	11	0.3	45.5%	54.5%
IT	51	13	0.3	69.2%	30.8%
Agric&Foods	5	1	0.2	0.0%	100.0%
Telecom	52	8	0.2	75.0%	25.0%
ConsGoods	22	2	0.1	100.0%	0.0%
Handl/Printing	27	1	0.0	0.0%	100.0%
Agric&FoodProcess-Machines	1	0	0.0	-	-
ConstrTechn	2	0	0.0	-	-
MachineTools	14	0	0.0	-	-
Motors	31	0	0.0	-	-
NuclearTechn	2	0	0.0	-	-
ThermProcesses	8	0	0.0	-	-
Transportation	45	0	0.0	-	-
Total	843	1077	1.3	55.2%	44.8%

Figure 3 illustrates the distribution of the responses to the question on the impact of scientific literature on the conception or implementation of R&D for the invention. The survey asked respondents the following question: "If it had not been for the scientific literatures published in the 15 years preceding the invention, how would have been the conception or implementation of the R&D for your invention affected?"[2]

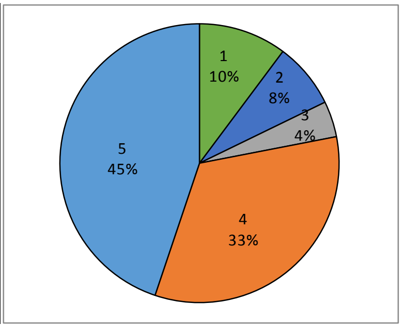
For 10.2% of inventions the idea for the invention could not have been conceived if it were not for the scientific literature. Moreover, 7.6% of inventors answered that the implementation of R&D would have been extremely difficult. This indicates that the scientific literature was an essential knowledge source for approximately 18% of inventions. Furthermore, we find that additionally for 4.2% of the R&Ds, the results

[2] The same question was asked with respect to the research equipment or material, and the collaboration with university/public research institute. The Appendix provides a summary result. It also presents an estimation result assessing how higher scientific absorptive capacity and higher risk preference of inventors promote the exploitation of scientific knowledge in their inventions.

would have been significantly delayed. Therefore, for approximately 22% of inventions, scientific literature was an important knowledge source for the conception or implementation of the R&D.

Figure 3. Impact of the scientific literature

		If it had not been for the scientific literature published in the 15 years preceding the invention,	frequency	share
influential / important / essential	1	the idea for the invention could not have been conceived	86	10.2%
influential / important / essential	2	implementation of R&D would have been extremely difficult	64	7.6%
influential / important	3	getting a result would have been significantly delayed	35	4.2%
influential	4	getting a result would have been slightly delayed	280	33.2%
No impact	5	it would have had no impact	378	44.8%

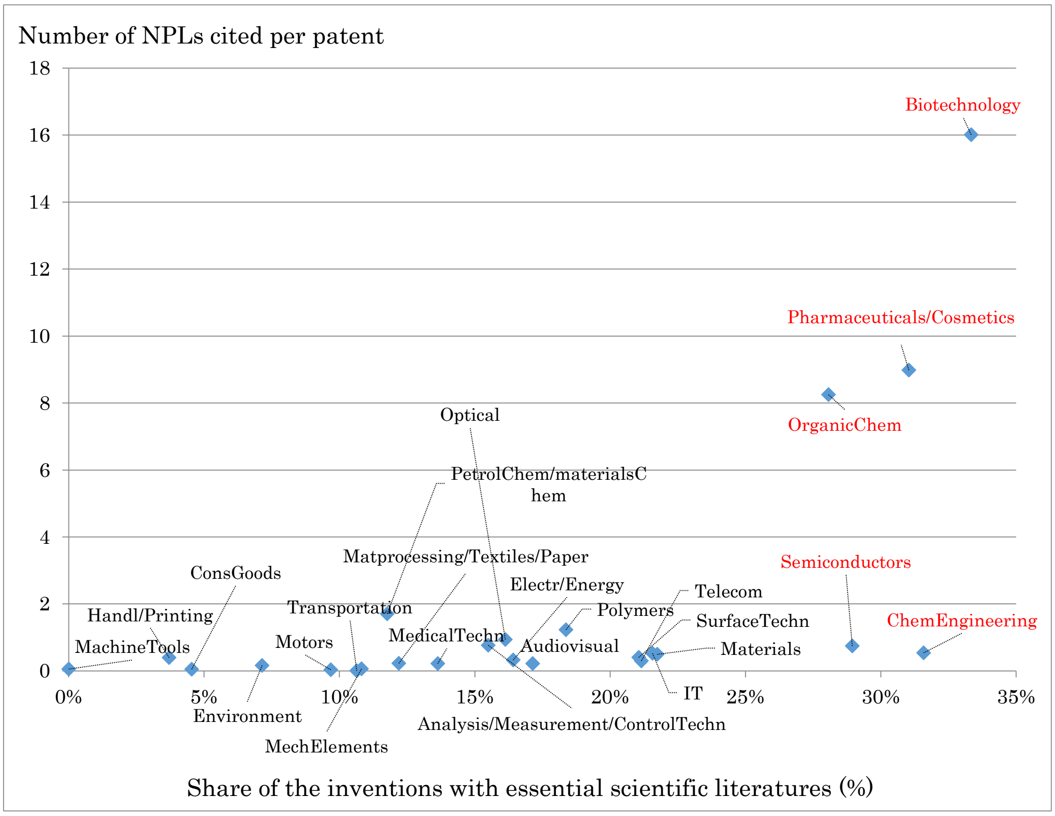


4. Completeness and noise of the citation information

4.1 Sector level evidence

Figure 4 shows the correlations between the average number of NPLs cited per patent and the share of the inventions that have an essential scientific literature across technology fields. It is clear that the average number of NPLs cited poorly predict the essentiality of science for the inventions. In particular, it is almost zero for a large number of technology fields where a substantial share of the inventions has an essential scientific literature. In particular, science literatures are frequently essential for Chemical Engineering and Semiconductors but the NPL citations are very low in these two fields. Thus, patent citations are substantially incomplete for many fields, including these two sectors. Only three technology fields (Biotechnology, Pharmaceuticals/ Cosmetics and Organic Chemistry) have a high level of NPL citations per patent application, and science literature are often essential for such fields. Excluding them, the average number of NLPs does not rise with the share of the inventions with essential science literature as knowledge source. This indicates the possibility that NPLs are often noise for predicting the existence of the important science literature as knowledge source.

Figure 4. Essentiality of science for inventions and frequency of NPL citations across sectors



4.2 Patent level evidence

Figure 5 presents a summary assessment of the completeness and the noise of the citation information, based on the patent level evidence matched with survey. In this figure, we compare the share of the inventions that have important scientific literature ("Inventions with important scientific literature sources") between the following two groups: the inventions that cite the non-patent literatures in the patent documents (N=176) and the inventions that do not cite any non-patent literature (N=667). We measure the completeness of the revealed citations by $\frac{C}{A+B+C}$, and the level of noise by $\frac{B+C}{B+C+D}$. The group of 36 inventions labelled as "B" (ambiguous cases) have incomplete references either in the survey response or in the patent documents; they cite NPLs but cannot be clearly matched to the survey responses.

As for the completeness, among the 185 inventions with important scientific sources (A+B+C), only 32 inventions (17%) reveal such important scientific literatures in an identifiable manner in their patent documents. To put it the other way, even if the inventions actually have important scientific sources for the conception and/or the implementation of the R&D, 83% of those inventions do not reveal the scientific sources,. Even if we include the ambiguous cases (B) in the numerator ($\frac{B+C}{A+B+C}$), only 37% of inventions reveal scientific sources in the patent documents. Thus, the patent citation information is quite incomplete.

As for the level of noise, among the 176 inventions citing any NPLs (B+C+D), 32 inventions (18%) have important scientific source corresponding exactly to the NPL cited. Therefore, 82% of the inventions with NPLs do not have important scientific sources. Even if we exclude the ambiguous cases (B) as noise, 61% of inventions have only noisy references in the patent documents. This indicates that much of the non-patent literatures revealed in the patent document are not the important scientific sources but the prior art only useful for clarifying the patentability or the other documents useful only for explaining the invention. Therefore, we can say that the citation data is a noisy index to trace the knowledge flow. However, we will see in Section 4 that the citation information provides still a useful information to identify whether there exists an important science source for the invention, even though it is incomplete and noisy.

Figure 5. Completeness and noise of the citation information

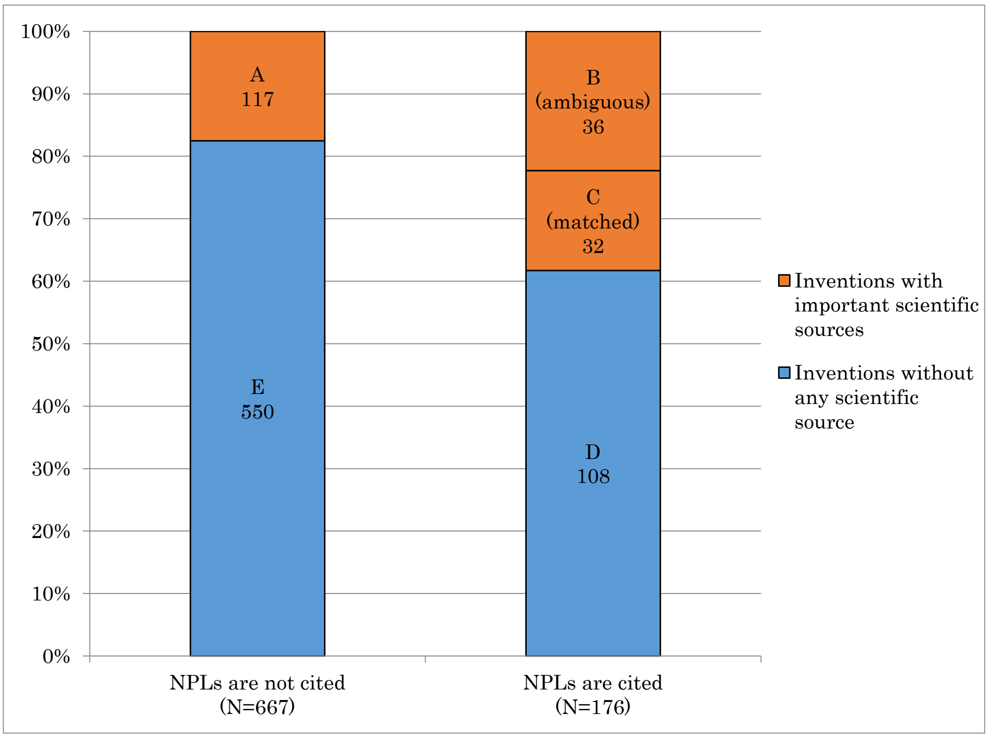


Table 2. Share of the inventions revealing scientific source

	N	%
Essential for conception	86	17.4%
Essential for implementation	64	15.6%
Important	35	20.0%
Total	185	17.3%

In Table 2, restricting the sample into the inventions that have important scientific sources, we compare the share of the inventions that reveal the important scientific literatures by essentiality of science for the inventions. In total, out of 185 inventions having important scientific sources, 17.3% inventions reveal those important scientific literatures in the patent documents. For the inventions with essential scientific source for conception (84 inventions), the share of the inventions revealing scientific source is 17.4%. For the inventions with an essential scientific source for the implementation of the R&D (64 inventions) the share is 15.6%, and for the inventions with not essential but important scientific source (35 inventions) the share is 20.0%. These results clearly suggest that the importance of literature as knowledge source for the invention does not have clear correlations with the probability of such literature being revealed in the patent document.

Table 3 shows the share of NPLs revealing the scientific sources by the place where the NPLs are referred to. In total, the share of the NPLs revealing scientific source is 7.6% (=82/1077). We can see that out of 82 NPLs revealing important scientific sources, 35 literatures (42.7%) are revealed at the place where the invention is described. This shows that the NPLs revealed as invention description provide important additional information to trace the knowledge flow, in addition to those revealed as prior art.

Table 3. Share of the scientific sources revealed by place (literature level)

		Total (literature level)		as prior art (literature level)		as invention description (literature level)	
		N	%	N	%	N	mean
Scientific source	Yes	82	7.6%	47	7.9%	35	7.3%
	No	995	92.4%	548	92.1%	447	92.7%
Total		1077	100.0%	595	100.0%	482	100.0%

5 Prediction model: Accuracy of citation information

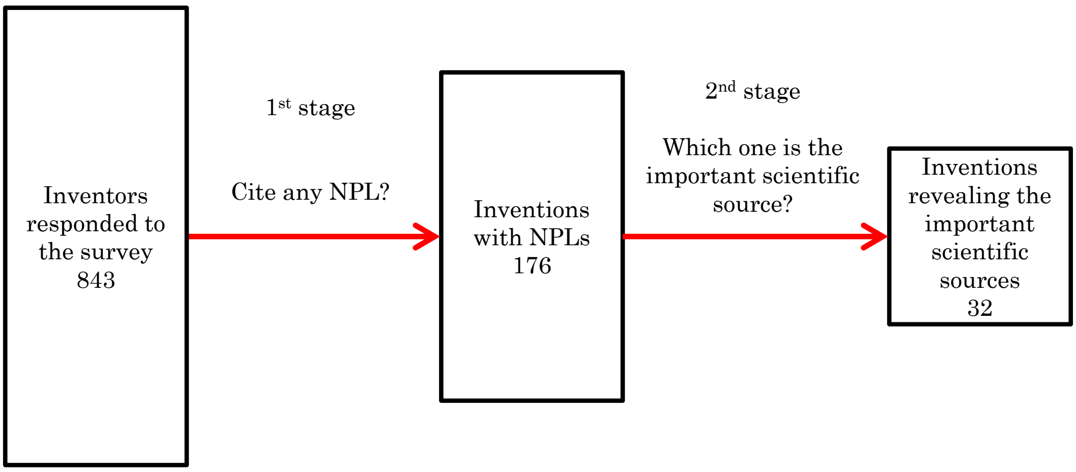
5.1 Specification for the prediction model and descriptive statistics

Given the above results, we estimate the "prediction model" that predicts the existence of important scientific source from the structure of the cited NPLs for that invention. We investigate whether there is any scientific source for invention when the invention reveals the NPLs in the patent document. This estimation helps us to understand the level of noise of citation information (measured by $\frac{B+C}{B+C+D}$).

Analytical framework and hypotheses

The specification of the prediction model is illustrated as Figure 6. In the first stage, we analyse the determinants of citing NPLs. Then, in the second stage, we identify the inventions with important scientific sources among the inventions revealing NPLs. This prediction model can provide a clue to discern the inventions with important knowledge sources from the citation data.

Figure 6. Prediction model



Given that there is competition for absorbing and exploiting recent scientific advances, early reference to an important scientific literature in an invention is likely to indicate that the inventor has exploited science for his invention. That is, if scientific literature which is highly cited by the other scientific papers is cited in a patent document soon after the publication of the scientific article, we can expect that such invention significantly uses the scientific knowledge embodied in the article. Our central hypothesis here is the following.

Hypotheses on the existence of important scientific source for the invention

If scientific literature with a larger number of forward citations is cited in a patent document early after the publication of the scientific article, such invention relies significantly on the science embodied in the article

Descriptive statistics

Table 4 compares the citation lag and the forward citation between the inventions that have important scientific literatures as knowledge sources and the inventions that do not have such literatures, restricting the sample to the inventions citing any NPLs[3]. We define the citation lag as the minimum length of the period between the priority date of the patent application and the publication years of the literature cited in that invention. Average forward citation in the Table 5 measures the aggregate average of the average number of forward citations of cited literatures of each focal invention. We take a logarithm of both the citation lag and the average forward citation, as these variables have skewed distribution. In Table 4, we find that the minimum citation lag is shorter for the inventions revealing important scientific literature. Moreover, the average number of forward citations is higher for the inventions with important scientific literatures. This suggests that the inventions which cite more recent scientific literature with large forward citations are more likely to have an important scientific source.

Table 4. Citation lag and forward citation (only for inventions with NPLs)

		Minimum citation lag (logarithm)		Average forward citation (logarithm)	
		N	mean	N	mean
Important scientific literature	Yes	67	1.61	44	4.81
	No	106	1.72	66	4.47
Total		173	1.67	110	4.61

[3] Data on the number of forward citation can be obtained only from Web of Science, while the citation lag can be calculated by using either Japanese literature database or Web of Science. Therefore, when we use the number of forward citation the sample is limited to the inventions citing literatures included in Web of Science.

Table 5. Prediction of the existence of scientific source

		Citation lag: Long	Citation lag: Short	Total
Forward citation	High	12.9% (N=31)	39.1% (N=23)	24.1% (N=54)
	Low	12.5% (N=24)	23.1% (N=26)	18.0% (N=50)
Total		12.7% (N=55)	30.6% (N=49)	21.2% (N=104)

Focusing on the cross effect of citation lag and forward citation provides more clear view. Table 5 shows how the citation lag and the forward citation of the science literature cited by the focal patent can predict the existence of scientific source for such patent. In this table, the sample is limited to the inventions citing those NPLs, which are included in the Web of Science and have information on forward citations and publication year. We use the literature with the maximum forward citation if inventions cite multiple NPLs. We calculate the citation lag of the invention for the literature with maximum forward citation. Restricting our sample to the inventions citing NPLs, we classified the inventions based on the median of the average forward citations of the cited literatures and the median of the average citation lag of the cited literatures with maximum forward citation. In Table 5, we find that among the inventions with highly cited literatures the share of the inventions with important scientific literature is 24.1%, while the corresponding share is 18.0% among the inventions with lower cited literatures. Similarly, the inventions with shorter minimum citation lag have higher probability of having important scientific literatures (30.6% vs. 12.7%). Especially, Table 6 shows that the share of inventions citing important scientific literatures is the highest for the inventions that have literatures with higher forward citation and shorter citation lag (39.1%). This suggests the possibility that the higher forward citation and shorter citation lag can predict the existence of scientific sources.

Estimation model

Specifically, we apply the following two-stage Heckman estimation models. There might be a selection bias in which the inventions with higher scientific capability have a greater propensity to rely on more recent scientific knowledge and to cite NPLs more frequently. To control for this bias, we estimate the following first stage model (1) and second stage model (2), using the instrumental variable; number of "patent" backward citations. We expect that propensity of citing patent literature has a correlation with the propensity of citing NPLs, while it does not have a correlation with the probability that the revealed NPLs are the actual scientific sources.

$$Cite_i = \alpha\ Scientific\ capability_i + \alpha_2 patent\ citation_i + \boldsymbol{\theta}\ Control\ variables_i + \epsilon_i. \quad (1)$$

$$\begin{aligned} Source_i = \beta_1\ Scientific\ capability_i + \beta_2 Citation\ lag_i + \beta_3 forward\ citation_i \\ + \beta_4 (Citation\ lag * forward\ citation)_i + \boldsymbol{\eta}\ Control\ variables_i + \varepsilon_i. \end{aligned} \quad (2)$$

In Equation (1), the dependent variable is the dummy variable indicating whether the inventor cites any NPLs (denoted by *Cite*). The independent variables $Scientific\ capability_i$ are the number of papers published by the inventor and by PhD degree. As an instrumental variable, we introduce the logarithm of the number of patent backward citations of the focal invention (denoted by *lncntpat*). The control variables are the number of inventors, the firm size, and the technology fields.

Equation (2) represents how we can identify the existence of important scientific sources for the invention, based on the structure of the revealed NPLs in the patent documents. The dependent variable is a dummy variable taking value 1 if the invention has important scientific sources.

For the independent variables, we introduce $Citation\ lag_i$ and $forward\ citation_i$, and the cross term of both variables $(Citation\ lag * forward\ citation)_i$ in addition to the $Scientific\ absorptive\ capacity_i$. By identifying the place where the literature is cited in the patent document, we differentiate the citation lag for the literatures cited as prior art and that for the literatures cited as description of the invention.

The descriptive statistics of the variables for the prediction model are shown in Table 6.

Table 6. Descriptive statistics of the variables in the prediction model

	Variable	Explanation	Obs	Mean	Std. Dev.	Min	Max
second stage	*Source*	dummy variable taking 1 if the inventions have important scientific sources	66	0.348	0.480	0	1
	fwcitation	logarithm of the maximum number of forward citations	66	4.59	1.93	1.10	9.25
	citationlag (full sample)	citation lag for the literatures with maximum foward citations	66	2.12	0.91	0.00	4.09
	*cross_fw*lag (full sample)*	cross term of fwcitation and citationlag	66	9.86	6.13	0.00	25.91
	citationlag (as prior art)	citation lag for the literatures with maximum foward citations disclosed as prior art	52	2.11	0.86	0.00	4.04
	*cross_fw*lag (as prior art)*	cross term of fwcitation and citationlag for the literatures disclosed as prior art	52	9.88	6.08	0.00	25.91
	citationlag (as description of invention)	citation lag for the literatures with maximum foward citations revealed for invention description	45	2.07	0.93	0.00	4.09
	*cross_fw*lag (as description of invention)*	cross term of fwcitation and citationlag for the literatures revealed for invention description	45	9.85	6.25	0.00	24.40
	selfcitation	dummy variable taking 1 if the invention has at least one self cited literature	66	0.11	0.31	0.00	1.00
	lnpaper	logarithm of the number of papers published by the inventors	66	2.06	1.37	0.00	5.25
	phd	PhD holder dummy	66	0.44	0.50	0.00	1.00
	lnnuminv	logarithm of the number of co-inventors	66	0.62	0.28	0.00	1.10
	lncntnpl	logarithm of the number of cited non-patent literatures	66	1.73	0.78	0.69	3.78
first stage	*Cite (full sample)*	dummy variable taking 1 if the inventions cite WOS literatures	621	0.185	0.389	0.00	1.00
	Cite (as prior art)	dummy variable taking 1 if the inventions cite WOS literatures at the place where the prior art is disclosed	619	0.147	0.354	0.00	1.00
	Cite (as description of invention)	dummy variable taking 1 if the inventions cite WOS literatures at the place where the invention is explained	492	0.130	0.337	0.00	1.00
	lnpaper	logarithm of the number of papers published by the inventors	621	1.065	1.162	0.00	6.22
	phd	PhD holder dummy	621	0.137	0.344	0.00	1.00
	lncntpat	logarithm of the number of cited patent applications	621	1.224	0.723	0.00	4.91
	lnnuminv	logarithm of the number of co-inventors	621	0.698	0.313	0.00	1.10

5.2. Estimation results for the prediction model: accuracy of citation

Table 7 shows the estimation results of Equations (1) and (2). To see the different effect depending on the places where the NPLs are cited, we divided our sample to the inventions with literature disclosed as prior art, the inventions with literatures cited as description of invention, and the full sample.

In the first-stage decision, we find that the instrumental variable, the number of patent backward citations (*lncntpat*), has a positive effect on the propensity of citing NPLs. Moreover, the results show that the scientific capability measured by the variable *lnpaper* has a positive effect.

The second-stage estimation results show that the number of forward citations has a positive effect. Furthermore, as hypothesized, the cross term of forward citation and citation lag have significantly negative signs. These results suggest that we can partially predict the existence of actual scientific sources by focusing on the number of forward citations and the citation lag of the revealed literatures in the patent documents. Especially, the inventions that have NPLs with higher number of forward citations and shorter citation lag are more likely to have scientific sources. In summary, the literature disclosed as prior art and revealed as invention description in the patent document can provide useful information to trace the knowledge flow even though the citation information has much noise[4].

We investigate the determinants of using scientific knowledge in Appendix and show the importance of scientific absorptive capacity and risk preferences of inventors to apply scientific knowledge to industrial invention.

Table 7. Estimation results for the prediction model

	second stage Source				first stage Cite			
	all		cited as invention description	cited as prior art	all		cited as invention description	cited as prior art
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
fwcitation	0.224***	0.232***	0.240***	0.270***				
	(3.06)	(3.10)	(2.91)	(3.80)				
citationlag	0.331**	0.346**	0.376**	0.316**				
	(2.34)	(2.40)	(2.29)	(2.28)				
*cross_fw*lag*	-0.098***	-0.099***	-0.111***	-0.103***				
	(-3.14)	(-3.09)	(-3.16)	(-3.31)				
selfcitation	0.416***	0.415***	0.365**	0.361***				
	(2.74)	(2.66)	(2.15)	(2.77)				
phd	-0.187		-0.273	-0.351***	0.320		0.336	0.259
	(-1.40)		(-1.53)	(-2.63)	(1.30)		(1.36)	(0.93)
lnpaper	0.044	0.018	0.045	0.043	0.304***	0.366***	0.314***	0.293***
	(0.79)	(0.29)	(0.56)	(0.77)	(3.66)	(5.31)	(3.66)	(3.12)
lncntpat					0.354***	0.352***	0.212*	0.409***
					(2.99)	(3.00)	(1.83)	(3.16)
lnnuminv	-0.046	-0.099	0.093	-0.312*	-0.169	-0.146	-0.092	-0.312
	(-0.27)	(-0.56)	(0.44)	(-1.66)	(-0.60)	(-0.52)	(-0.31)	(-0.92)
lncntnpl	-0.054	-0.043	-0.063	-0.108*				
	(-0.88)	(-0.69)	(-0.82)	(-1.78)				
Constant	-0.288	-0.487	-0.217	0.418	-1.888**	-1.919**	-1.872**	-2.555***
	(-0.42)	(-0.69)	(-0.22)	(0.63)	(-2.19)	(-2.25)	(-2.16)	(-3.05)
firmsize	yes	yes	yes	yes	yes	yes	yes	yes
mainarea	yes	yes	yes	yes	yes	yes	yes	yes
Observations	66	66	52	45	587	588	597	617
Censered	521	522	545	572	521	522	545	572
rho	0.398	0.506	0.279	-0.538	0.398	0.506	0.279	-0.538
chi2	32.71	28.37	32.51	49.70	32.71	28.37	32.51	49.70

z-statistics in parentheses
*** p<0.01, ** p<0.05, * p<0.1

6. Conclusion

This study assessed the completeness and the level of noise of inventor citation information at the patent level, using an original inventor survey on scientific sources and the NPLs dataset from *purely inventor citations*. We directly identified the actual knowledge sources of the inventions, and compared them to the revealed NPLs in the patent documents. The survey results indicate that science significantly contributes to the Japanese invention activities. Over a half of inventions rely on scientific literature

[4] Our sample includes the researchers working at university (42 inventors out of 843 inventors). However, the results are robust even if we limit the sample into only the corporate inventions.

in their R&D process. Especially, for approximately 18% of inventions, the scientific literature was essential for the conception and implementation of R&D. We also find that higher scientific absorptive capacity and higher risk preferences of inventors promotes the exploitation of the scientific knowledge to the inventions.

We find only 17% of the inventions that have important scientific source reveal such important literature in the patent documents. Even if we include the ambiguous cases due to incomplete survey responses and/or incomplete references in the patent documents, the share of such inventions is 37%. Therefore, citation information is significantly incomplete. Moreover, more important science sources are not necessarily more revealed. We also find that the NPLs revealed at the place where an invention is described provide important additional information in identifying the science sources (around 40%).

At the same time, even if the inventions reveal NPLs in their patent documents, 82% of those inventions do not have any important scientific sources for the conception or for the implementation of the invention. When we do not treat the above ambiguous cases as noises but treat as matched cases, the share of inventions citing only unimportant NPLs is reduced but still amounts to 61%. These results indicate that citation information is also a noisy index.

However, our estimation results show that citation information is still a useful index to trace knowledge flow. Controlling for the selection of citing NPLs in the patent document, we find that the inventions are more likely to have scientific sources when they cite NPLs with higher number of forward citations and shorter citation lag. This result was observed both for the inventions disclosing NPLs as prior art and for the inventions revealing NPLs for describing invention. These findings indicate that the citation data, both disclosed as prior art and revealed as invention description, helps us in identifying important knowledge sources, though it is very incomplete and noisy.

We can draw the following policy implications and future research agenda. First, patent backward citations, even if extended to those made in the part describing the invention, still significantly underestimate the contributions of science to corporate inventions. Science plays an essential role for a corporate R&D even if the patent does not refer any scientific and technical literature. A direct survey instrument would be essential for correctly assessing the contribution of science. Second, references made in the part describing the invention (examples implementing the invention, its utility etc.) are important additional source for identifying the knowledge source. The citation lag and the forward citations of the scientific literatures provide important clues for such identifications. Third, the significant sectoral variations in the relation between the existence of essential science source for inventions and the NPL citations remain a puzzle, waiting for a further study.

References

Alcacer, J. and M. Gittelman (2006) "Patent Citations as a Measure of Knowledge Flows: The Influence of Examiner Citations," Review of Economics and Statistics, vol. 88, pp.774-779.

Alcacer, J., Gittelman, M., & Sampat, B. (2009). Applicant and examiner citations in US patents: An overview and analysis. Research Policy, 38(2), 415-427.

Criscuolo, P. and B. Verspagen (2008) "Does It Matter Where Patent Citations Come From? Inventor vs. Examiner Citations in European Patents," Research Policy 37, pp.1892-1908.

Duguet, E. and M. MacGarvie (2005) "How Well Do Patent citations Measure Flows of Technology? Evidence from French Innovation Surveys," Economics of Innovation and New Technologies 14, pp.375-394.

Jaffe, A., M. Fogarty, and B. Banks (1998) "Evidence from Patents and Patent Citations on the Impact of NASA and Other Federal Labs on Commercial Innovation," Journal of Industrial Economics, 46(2), pp.183-205.

Jaffe, A., M. Trajtenberg, and M. Fogarty (2000) "Knowledge Spillovers and Patent Citations: Evidence from a Survey of Inventors," American Economic Review, vol. 90, pp.215-218.

Klevorick, A., R. Levin, R. Nelson, and S. Winter (1995) "On the Sources and Significance of Inter-industry Differences in technological Opportunities," Research Policy 24, pp.185-205

Mansfield E. (1995) "Academic Research Underlying Industrial Innovations: Sources, Characteristics, and Financing", Review of Economics and Statistics, Vol. 77, pp. 55-65

Nagaoka, S., N. Tsukada, K. Onishi and Y. Nishimura (2012) "Innovation Process in Japan in the Early 2000s as Seen from Inventors: Agenda for Strengthening Innovative Capability", RIETI Discussion Paper Series 12-J-033 [in Japanese]

Nelson, A. (2009) "Measuring Knowledge Spillovers: What Patents, Licenses and Publications Reveal about Innovation Diffusion," Research Policy, vol. 38, pp.994-1005.

Roach, M. and W. Cohen (2013) “Lens or Prism? Patent Citations as a Measure of Knowledge Flows from Public Research”, Management Science, Vol. 59, No. 2, pp. 504–525.

Appendix 1. Research equipment/material and collaboration with university/public research institute as knowledge sources for inventions

Figure A.1 and A.2 show the impacts of research equipment/material and collaboration with university/public research institute. In Figure A.1, we find that research equipment or material is essential for 16.3% of inventions. Moreover, the figure shows that 21.0% of inventors perceived the research equipment or material as important in their R&D process. Figure A.2 shows that for 2.7% of corporate inventions, collaboration with university/public research institute is essential for conceiving or implementing R&D project. We also find that collaboration with university/research institute has important influence on 6.9% of inventions.

Figure A.1 Impact of the research equipment or material

		If it had not been for the scientific knowledge embodied in research materials or equipment in the 15 years preceding the invention,	frequency	share
influential / important / essential	1	the idea for the invention could not have been conceived	71	7.8%
influential / important / essential	2	implementation of R&D would have been extremely difficult	72	8.5%
influential / important	3	getting a result would have been significantly delayed	39	4.7%
influential	4	getting a result would have been slightly delayed	258	30.6%
No impact	5	it would have had no impact	403	48.4%

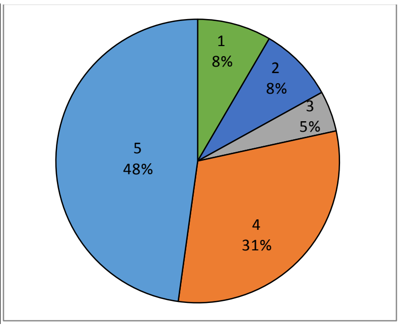
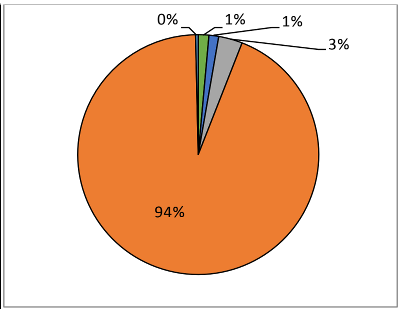


Figure A.2 Impact of the collaboration with university/public research institute

		If it had not been for the collaboration with university or public research institute in the years preceding the invention,	frequency	share
influential / important / essential	1	the R&D project would have not been conceived because of the lack of seeds or needs	12	1.4%
influential / important / essential	2	implementation of R&D would have been terminated	11	1.3%
influential / important	3	internal R&D would result in the significant reduction of the R&D size or the significant delay in getting a result	27	3.2%
No impact	4	it would not have a significant impact	790	93.7%



Appendix 2. Determinants of the use of science

This appendix examines the determinants of the use of science. In Table A1, we compare the share of the inventions that use the scientific knowledge embodied in literature, research equipment/material and university researchers, focusing on the absorptive capacity measured by the possession of PhD degree. We regard the invention that has any scientific source affecting R&D as the invention using scientific knowledge. We expect that inventors with higher absorptive scientific capacity are more likely to rely on the scientific knowledge in their invention process.

Table A1 shows that the inventors with PhD degree have higher rate of using scientific knowledge in their invention process. For example, 66.4% of inventions developed by PhD holders use the scientific literature, while only 50.4% of inventors without PhD degree use the scientific literature for their inventions. Similarly, research equipment/material is used as a scientific source by 61.1% of the inventors with PhD degree, and by 50.5% of the inventors without PhD. Moreover, direct collaboration with university has important influence on R&D for 12.2% of the inventors with PhD degree, while it has importance on only 4.8% of the inventors without PhD. These results indicate the importance of scientific absorptive capacity on the usage of scientific knowledge.

Table A1. Usage rate of science by the possession of PhD

	Use rate		
	literature	equipment material	university
Yes (N=131)	66.4%	61.1%	12.2%
No (N=711)	50.4%	50.5%	4.8%
Total (N=842)	55.9%	52.1%	5.9%

Specification

Our research question here is how the probability that the invention has any scientific source (scientific literature, research equipment/material, or direct collaboration with university researchers) varies, depending on the scientific absorptive capacity of the inventor and the firm as well as on the risk preference of the inventor. Specifically, we estimate the following model, identifying separately the effect on the use of three channels:

$$Use_i = \alpha_1\, Scientific\ absorptive\ capacity_i + \alpha_2 Complementary\ asset\ for\ science_i + \alpha_3 Risk\ loving_i + \boldsymbol{\theta}\ Control\ variables_i + \epsilon_i. \quad (3)$$

In this specification, i denotes the invention and the vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ are the coefficient parameters. The dependent variable is the dummy variable that takes a value of 1 if the scientific literature, research equipment/material or university has some influence for the conception and/or implementation of R&D. For the independent variables, we measure the $Scientific\ absorptive\ capacity_i$ by the number of papers published by the inventors (denoted by *lnpaper*) and the Ph.D. holder dummy (denoted by *phd*). The variable $Complementary\ asset\ for\ science_i$ is measured by the firm's equipment for R&D (*equip*) and the environment for technological development (*techenv*). These indices are the responses to the previous inventor survey, which was given on a five-point scale. The variable $Risk\ loving_i$ is also the response to the previous survey on the willingness to take risks, with answers given on an eleven point scale (*riskloving*). We expect that these indices affect positively the decision of whether to use science.

As the control variables, we include the number of inventors which measures the input of R&D as well as the type of R&D&; the firm size which is related to the application strategy; and the technology fields. For firm size, we use the dummy variables taking the value of 1 if the number of employees is 1,000 or more and less than 5,000 (*emp1000*), and the number is more than 5,000 (*emp5000*). Moreover, we control for the difference of technology fields. The descriptive statistics of these variables are shown in Table A2.

Table A2. Descriptive statistics for identifying the determinants of using science

	Variable	Explanation	Obs	Mean	Std. Dev.	Min	Max
literature	*Use*	dummy variable taking 1 if the scientific literature has influence for the conception and/or implementation of R&D	583	0.575	0.495	0	1
	phd	PhD holder dummy	583	0.139	0.346	0	1
	lnpaper	logarithm of the number of papers published by the inventors	583	1.093	1.161	0	6.22
	riskloving	willingness to take risks	583	6.583	2.475	1	11
	equipment	completeness of the equipment for R&D	583	3.631	1.163	1	5
	techenv	environment for technological development	583	3.247	1.091	1	5
	lnnuminv	logarithm of the number of co-inventors	583	0.688	0.310	0.00	1.10
material/ equipment	*Use*	dummy variable taking 1 if the scientific material/equipment has influence for the conception and/or implementation of R&D	585	0.542	0.499	0	1
	phd	PhD holder dummy	585	0.142	0.349	0	1
	lnpaper	logarithm of the number of papers published by the inventors	585	1.101	1.169	0	6.22
	riskloving	willingness to take risks	585	6.593	2.476	1	11
	equipment	completeness of the equipment for R&D	585	3.634	1.163	1	5.00
	techenv	environment for technological development	585	3.248	1.093	1	5
	lnnuminv	logarithm of the number of co-inventors	585	0.688	0.309	0.00	1.10
university	*Use*	dummy variable taking 1 if the scientific literature has influence for the conception and/or implementation of R&D	561	0.066	0.248	0	1
	phd	PhD holder dummy	561	0.148	0.355	0	1
	lnpaper	logarithm of the number of papers published by the inventors	551	1.121	1.174	0	6.22
	riskloving	willingness to take risks	561	6.560	2.488	1	11
	equipment	completeness of the equipment for R&D	561	3.631	1.167	1	5
	techenv	environment for technological development	561	3.253	1.076	1	5
	lnnuminv	logarithm of the number of co-inventors	561	0.686	0.308	0.00	1.10

Estimation results for the determinants of using science

Table A3 shows the estimation results of specification (3) with probit estimation. As for the use of scientific literature, the coefficients of the number of papers and willingness to take risks are positive and significant. For the use of research equipment/material, the variable *phd*, instead of *lnpaper*, and *riskloving* have statistical significant effects. For the importance of collaboration with university, both *phd* and *lnpaper* have positive effect. The variable *riskloving* also has positive effect on the use of university. These results suggest that high scientific absorptive capacity and high risk preferences of inventors promotes the exploitation of the scientific knowledge to the corporate inventions. Especially, our finding indicates the importance of PhD holders to apply advanced research equipment and to collaborate with university.

Table A3. Determinants of using scientific knowledge

	Use								
	literature			material/equipment			university		
	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
phd	0.004		0.085	0.145*		0.157**	0.034		0.074***
	(0.05)		(1.34)	(1.96)		(2.49)	(1.27)		(2.84)
lnpaper	0.054**	0.053***		0.009	0.027		0.009	0.013**	
	(2.33)	(2.64)		(0.38)	(1.37)		(1.20)	(2.26)	
riskloving	0.016*	0.018**	0.023***	0.011	0.013	0.014*	0.005*	0.006*	0.006*
	(1.84)	(2.15)	(2.70)	(1.24)	(1.51)	(1.71)	(1.69)	(1.74)	(1.81)
equip	-0.024	-0.018		-0.003	-0.005		-0.003	-0.003	
	(-1.15)	(-0.97)		(-0.14)	(-0.27)		(-0.41)	(-0.54)	
techenv	0.017		0.016	-0.005		-0.004	-0.003		-0.004
	(0.74)		(0.80)	(-0.21)		(-0.22)	(-0.38)		(-0.51)
lnnuminv	-0.024	-0.033	-0.028	-0.025	-0.023	-0.040	-0.067***	-0.064***	-0.065***
	(-0.35)	(-0.48)	(-0.42)	(-0.36)	(-0.34)	(-0.60)	(-2.97)	(-2.89)	(-2.84)
firmsize	yes	yes	yes	yes	yes	yes	yes	yes	yes
mainarea	yes	yes	yes	yes	yes	yes	yes	yes	yes
Observations	583	591	617	585	593	619	551	558	585
Pseudo R2	0.0578	0.0571	0.0502	0.0356	0.0296	0.0368	0.135	0.130	0.124

z-statistics in parentheses
*** p<0.01, ** p<0.05, * p<0.1
coefficients are marginal effects