

The survey of industrial R&D—patent database link project

William R. Kerr · Shihe Fu

Published online: 26 January 2008
© Springer Science+Business Media, LLC 2008

Abstract This article details the construction of a firm-year panel dataset combining the NBER patent dataset with the Survey of Industrial R&D conducted by the Census Bureau and National Science Foundation. The dataset constitutes a platform that offers an unprecedented view of the R&D-to-patenting innovation process and a close analysis of the strengths and limitations of the R&D survey. The files are linked through a name-matching algorithm customized for uniting the firm names to which patents are assigned with the firm names in the Census Bureau's SSEL business registry. Through the Census Bureau's file structure, R&D can be linked to the operating performances of each firm's establishments, further facilitating innovation-to-productivity studies.

Keywords Innovation · Research and development · Patents · Scientists · Technology

JEL Classifications C81 · O30 · O31

1 Introduction

The development and diffusion of new innovations are central to economic growth. In many theoretical models, such as the textbook Solow and Swan framework, this technology progress is the central driver of long-run productivity gains and higher standards of living. As the majority of these investments are undertaken by the private sector in the US, understanding the firm-level underpinnings of technology progress is important to

W. R. Kerr (✉)
Harvard Business School, Boston, MA 02163, USA
e-mail: wkerr@hbs.edu

S. Fu
Research Institute of Economics and Management, Southwestern University of Finance
and Economics, Chengdu 610074, P.R. China
e-mail: fush@swufe.edu.cn

academics, policy makers, and business managers. Essential questions include which firms invest in research and development (R&D), how their resulting innovations spread to other firms, how technology adoptions are translated into within-firm operating gains, and how productivity growth at the firm level aggregates to overall economic performance (including reallocations across firms).

A thorough understanding of these microeconomic phenomena promises to inform better management practices and policy prescriptions. Not surprisingly then, many empirical researchers have examined, either jointly or separately, corporate R&D and technology diffusion. This empirical work has confronted, however, significant data constraints. The first constraint is the R&D information collected for individual firms. Compustat, a commercial database developed by Standard & Poor's, is widely employed.¹ This sample, however, is restricted by construction to publicly listed firms only. Further, firm-level aggregates ignore the importance of line-of-businesses within firms (e.g., Levin et al. 1985; Cohen et al. 1987) or the within-firm location choices of R&D (e.g., Adams and Jaffe 1994). They also do a poor job of linking R&D efforts to the capital investments by firms, the opening or closure of operating plants and product lines, and so on. Research over the past two decades employing the micro-records of firms repeatedly stresses the importance of the tremendous heterogeneity that exists (e.g., Davis et al. 1996).

Empirical studies of technology diffusion, on the other hand, often start with the micro-records of individual patents. The NBER patent dataset, originally compiled by Hall et al. (2001), offers some unique advantages. It affords a comprehensive view of US patenting that includes public and private firms, universities, and government laboratories. Unlike the Compustat R&D data, the micro-records of patents also allow firm-level patenting to be disaggregated by technologies, inventor locations, or both. Moreover, the citation patterns across these patents offer a tractable view of inventor-to-inventor communications within and across firms (e.g., Jaffe et al. 2000). Too often, however, these diffusion studies start and end with analyses of citations. While citations are informative, a deeper study of technology growth and diffusion should link these knowledge transfers to the actual economic outcomes for firms. The disaggregated patent data and citation flows will realize their full potential only when paired with disaggregated R&D investments and operating outcomes within firms.

The core methodological innovation of the dataset described in this article is the matching of external patent files to the Census Bureau data family. The dataset encompasses the National Science Foundation's Survey of Industrial Research and Development, the NBER patent dataset, and the Census Bureau's establishment-level operating data. The dataset constitutes a platform that offers an unprecedented view of the R&D-to-patenting innovation process.

The backbone for this platform is the Census Bureau's firm-level linkage of the Survey of Industrial Research and Development (RAD) to the plant-level operating data collected in the economic surveys of the Census Bureau. The latter include annual employments and wages, as well as industry and geographic codes, for all private-sector establishments in the US. Moreover, the Census Bureau file structure facilitates incorporating richer establishment-level characteristics (e.g., investments and outputs) collected in sector-specific surveys like the Census of Manufacturers. Jarmin and Miranda (2002) and Davis et al. (1996) describe in detail the Longitudinal Business Database and Census of Manufacturers, respectively.

¹ For more information see <http://www.compustat.com>.

Section 2 provides an overview of the RAD and the development of a core panel of major R&D firms that are closely monitored. This discussion should help researchers interested in corporate innovation to understand the RAD's major advantages and limitations for empirical work. The Census Bureau and National Science Foundation (NSF) are both encouraging research proposals that employ the RAD.

Section 3 describes the NBER patent dataset (PAT) and the RAD–PAT matching process. The PAT records are matched into the Census Bureau's data through firm names. This effort concentrates on the core RAD panel developed in Sect. 2, matching over 90% of the large R&D firms and manually verifying the unmatched records. In total, approximately 85% and 70% of US corporate R&D expenditures and patenting, respectively, are appropriately linked. The resulting dataset includes the most detailed and disaggregated information available on business R&D expenditures, patenting, and operating activities. The dataset covers over 30 years and includes all major firms investing in R&D and patenting in the US—be they public or private, US-owned or foreign-owned, and so on.

Section 4 describes how patents and R&D expenditures are allocated within firms across states and industries. The last section concludes by describing in greater detail potential research applications.

A companion working paper (Kerr and Fu 2006a) provides significantly more detail around the RAD and PAT datasets employed, the name-matching procedures developed for pairing firms, the panels of R&D firms constructed, and so on. This documentation discusses individual firm data, however, and is therefore restricted to researchers who have obtained appropriate security clearances through the Census Bureau.²

2 Survey of industrial research and development

The RAD is the US government's primary instrument for surveying the R&D expenditures and innovative efforts of US firms. This is an annual survey conducted jointly by the Census Bureau and NSF. The survey includes both public and private firms, as well as foreign-owned firms, undertaking R&D within the United States. The information collected from this survey is aggregated for publications like *Science and Engineering Indicators*, *National Patterns of R&D Resources*, and *R&D in Industry*.³

With appropriate clearance, researchers can access the base RAD survey responses through the Census Bureau. These micro-records span 1972–2000 and provide the most detailed statistics available on firm-level R&D efforts; moreover, the records can be linked to the Census Bureau's firm-level operating data for rich empirical analyses of the output and productivity gains from these investments. This section begins with a description of the RAD's core variables and the construction of a sample panel of major R&D producers. The section closes by outlining supplementary R&D panels that can be developed for the largest R&D performers and by discussing empirical strategies.

Adams and Peck (1994) provide a rich history of the RAD and changes in the RAD's sampling frame over time. Each year, the RAD surveys with certainty the identified firms that are conducting R&D within the US over a nominal expenditure bar. This expenditure hurdle began at \$500 k in the 1970s, was raised to \$1 m for most of the 1980s and 1990s, was raised again to \$5 m in 1996, and was most recently adjusted to \$3 m in 2002. While firms undertaking less than this bar are sub-sampled, these records are more difficult to

² For more information see <http://www.websserver02.ces.census.gov/index.php/ces/1.00/researchguidelines>.

³ Available at <http://www.nsf.gov/statistics>.

employ due to their uneven coverage, unbalanced panel, and frequently imputed values. The Census Bureau and NSF are constantly working to update the sample frame to include new R&D performers although identification lags can occur. The strength of the RAD is in the analysis of the major R&D firms that are repeatedly observed and comprise the bulk of US R&D.

Surveyed firms are legally required to provide five mandatory items:

- Total costs incurred for R&D within the firm (RDTOT).
- Domestic net sales and receipts of the firm (DNS).
- Domestic net employment of the firm (DNE).
- Federally funded R&D performed within the firm (RDFED).
- State location of R&D performance (added in 2002).

Additional information is also requested on a voluntary basis. Three optional questions are frequently asked and of significant interest to researchers studying innovation and technology transfer:

- Number of R&D scientists and engineers (SET).
- Total company funds for R&D activities financed by the company but performed by others outside the company within the US (OUTUSCOMP).
- Total company funds for R&D activities performed by foreign subsidiaries or by other organizations outside the US (OUTFOREIGN).

The core variable RDTOT measures domestic, within-firm R&D expenditures. RDTOT includes R&D supported by US federal funds undertaken by the firm, but it excludes all foreign-sourced R&D efforts (OUTFOREIGN) or US R&D efforts undertaken outside of the firm (OUTUSCOMP). These latter two variables are important for full descriptions of R&D efforts in industries where a substantial fraction of R&D is outsourced or conducted overseas (e.g., pharmaceuticals).

The first column of Table 1 lists the total number of RAD observations with positive RDTOT after a small number of duplicate and subsidiary records are culled. In the case of subsidiary records, it was confirmed that parent company records combine all of the information contained in the subsidiary observations. The base RAD files contain significantly more observations in the 1990s than in earlier decades due to changes in record retention. Surveyed firms that reported no R&D are dropped from the data prior to 1992, while all firm records are retained today. The reported observation counts for firms with positive R&D are more stable. Throughout the 1970s and 1980s, approximately 3,000 firms are included annually. This firm count rises in the early 1990s before retreating by the close of the decade. This surge and decline reflects changes in the nominal expenditure hurdle and adjustments to industry sampling procedures.

Table 1 continues with the unweighted and weighted sums of the mandatory variables US R&D (RDTOT), US sales, and US employment and the voluntary variable scientists and engineers. Expenditures are in nominal dollars. The sums for the scientists, sales, and employment variables are calculated over the firms reporting positive RDTOT. The raw aggregates provide a baseline for comparing the R&D expenditure incorporated in the balanced panel discussed next. The weighted nominal aggregates, on the other hand, afford a comparison to the published NSF statistics for the US. As would be expected, the RAD sums closely mirror the published data, with similar levels and highly correlated growth patterns. The minor differences can arise due to post-processing edits or revised weights. The trends for the summed counts of scientists and engineers also align with aggregate science and engineering employment estimates

Table 1 Mandatory variables summary statistics

Year	Observations with positive US R&D	Unweighted RAD sums				Weighted RAD sums			
		US R&D (\$b)	Sci./ Eng. (k)	US sales (\$b)	US empl. (m)	US R&D (\$b)	Sci./ Eng. (k)	US sales (\$b)	US empl. (m)
1974	3,233	22.3	345	741	13.8	23.1	364	776	14.4
1975	3,053	23.6	348	736	13.3	24.1	364	769	13.9
1976	3,082	26.1	368	800	13.5	26.6	384	839	14.2
1977	3,042	29.2	380	838	13.5	29.5	391	869	14.1
1978	2,980	32.8	404	927	14.5	33.4	418	962	15.1
1979	2,986	37.2	443	1,357	15.2	37.9	457	1,396	16.0
1980	2,968	42.9	451	1,458	16.7	43.8	468	1,505	17.6
1981	3,049	50.5	482	1,675	15.9	51.8	511	1,766	17.1
1982	2,982	57.6	504	1,654	14.8	58.9	533	1,718	15.9
1983	2,595	58.7	495	1,613	13.4	60.3	520	1,692	14.0
1984	2,597	68.9	521	1,829	14.1	71.0	550	1,916	14.8
1985	2,579	76.0	546	1,851	14.0	78.2	575	1,938	14.6
1986	3,690	84.7	612	1,987	15.0	91.0	700	2,090	16.8
1987	3,737	89.3	620	2,097	15.2	96.4	720	2,220	17.4
1988	3,514	93.5	628	2,158	14.5	98.8	715	2,278	16.7
1989	3,399	95.4	612	2,294	14.2	101.6	655	2,439	16.5
1990	3,342	97.5	607	2,512	14.3	104.5	702	2,678	16.7
1991	3,299	95.5	646	2,386	13.5	102.4	728	2,557	15.8
1992	5,028	105.6	659	2,836	15.1	121.8	778	3,068	16.7
1993	6,439	109.3	664	3,031	15.2	118.3	763	3,200	16.4
1994	4,883	110.8	648	3,294	15.3	119.6	749	3,594	17.4
1995	4,654	121.9	715	3,429	15.0	132.0	833	3,918	17.7
1996	3,969	131.5	750	3,502	14.7	144.6	887	4,095	18.1
1997	3,741	139.9	782	3,698	14.6	157.5	951	4,571	20.2
1998	3,326	145.5	795	3,748	14.2	169.1	997	4,675	18.3
1999	3,671	153.6	814	4,111	14.2	182.7	1,033	5,841	22.9
2000	3,583	167.6	853	4,438	13.9	199.5	1,042	5,250	17.6

Notes: Raw and weighted summaries from RAD after basic culling of duplicated and divisional records. Sums for scientists and engineers, sales, and employment variables are calculated over observations with positive R&D. Expenditures are in nominal dollars. R&D totals do not include R&D performed outside of the US or R&D performed outside of the company within the US

derived from the Current Population Survey, the US Government's monthly survey of unemployment and labor force participation. These comparisons are available upon request.

Turning to firm-level analyses, it is critical to note that R&D investments are cumulative in nature, often modeled through stock metrics similar to investment and inventory. Moreover, there is a natural lag from when R&D investments are made to when operating benefits are realized. Thus, the RAD should not be simply linked at the firm-year level to other Census Bureau datasets for estimations; panel dataset techniques are instead required. The formation of panels where repeated observations on a firm's innovative investments

are captured allows for the cumulative history and lagged realizations to be accounted for appropriately. The important cost of this panel development, however, is that firms without a full history of R&D investments need to be excluded from the resulting panel data. Thus, the longer the panel constructed, the smaller the sample size of firms that can be appropriately incorporated.

Evaluating various panel dimensions and timeframes, the optimal span for the balanced panel constructed for this project is 1986–1996. A total of 1,333 firms with complete survey records from 1986–1996 and at least one year of positive RDTOT are selected. This initial draw does not exclude imputed RDTOT values, but does maintain the imputation flags for later pruning when required. Approximately 1,200 of these firms have positive R&D investment in at least 10 of the 11 years. This balanced panel represents about 80% of the total US R&D expenditures and is the backbone for the patent link effort described below.

This 11-year period makes the balanced R&D panel relevant for the 1987, 1992, and 1997 economic census years, when the plant-level operating data are most abundant, while also recognizing that the panel size diminishes as it is extended to earlier or later years. A stretched, unbalanced panel is also constructed. This unbalanced panel begins with the 1986–1996 balanced panel and further incorporates other observations for these firms from 1974–1985 or 1997–2000. This unbalanced panel dataset can be used to construct any other balanced panel across time periods that encompass the 1986–1996 period. This panel further serves as the foundation for extensions from 1997–2002 as the RAD files and economic censuses become available to researchers.

Table 2 reports similar summary statistics as Table 1 for the unbalanced panel. The middle section shows the 1,333 firms during the 1986–1996 balanced panel period, and the decline in the sample size in earlier and later years. Note that some firms may enter and leave during these additional years, depending on the firms' R&D activity and the sampling frame. The number of observations in an earlier year is thus an upper bound for the sample size that can be constructed.

Finally, large R&D performers are requested in the optional questions to disaggregate their R&D efforts on a number of dimensions. These optional variables were typically available in odd-numbered years only after 1977. The odd–even year collection pattern was dropped in 1998, with the optional variables now collected annually. Moreover, only the largest R&D producers regularly complete these detailed reports. Nevertheless, the optional questions provide unparalleled descriptions of the R&D efforts of major firms:

- By state for domestic within-firm R&D.
- By foreign country for foreign-sourced R&D.
- By basic field of science and/or applied technology field (discontinued in 1997).
- By federal agency sponsoring the within-firm R&D (e.g., NASA, Defense).
- By pollution abatement or energy type where applicable.

These detailed R&D descriptions are a powerful addition to the Census Bureau's establishment-level operating data. Adams and Jaffe (1994), for example, use this detail to study within-firm and across-firm R&D spillovers by geography and industry. Moreover, these snapshots offer a valuable foothold for designing instrumental variable specifications for firm-level R&D efforts that combine past firm-level R&D allocations with external trends (e.g., changes in federal funding initiatives).

In general, researchers will find the most traction with two empirical strategies that complement each other. The first approach focuses primarily on the mandatory R&D items and the three frequently reported optional questions noted above (i.e., SET,

Table 2 Mandatory variables summary statistics—balanced panel

Year	Total observations	Observations with positive US R&D	Means for observations with positive R&D				Unweighted sums for observations with positive R&D			
			US R&D (\$m)	Sci./Eng. (\$m)	US sales (\$m)	US empl. (k)	US R&D (\$b)	Sci./Eng. (k)	US sales (\$b)	US empl. (m)
1974	666	664	23.1	372	621	11.6	15.3	225	410	7.7
1975	695	695	26.1	394	650	11.8	18.1	252	450	8.2
1976	697	697	28.7	415	695	12.0	20.0	267	483	8.4
1977	685	685	32.4	445	739	12.4	22.2	273	504	8.4
1978	691	690	36.9	472	807	13.1	25.5	295	553	9.0
1979	697	697	41.2	506	1,240	13.6	28.7	324	862	9.4
1980	697	697	47.6	512	1,403	15.8	33.1	327	975	11.0
1981	807	807	47.3	518	1,386	12.4	38.2	355	1,100	9.8
1982	808	808	53.8	497	1,390	11.8	43.5	378	1,113	9.5
1983	820	810	55.9	478	1,454	11.2	45.2	378	1,176	9.1
1984	824	816	65.3	503	1,659	12.1	53.3	403	1,352	9.9
1985	825	815	73.0	533	1,709	12.2	59.5	428	1,391	9.9
1986	1,333	1,310	52.9	376	1,166	8.4	69.3	492	1,527	11.0
1987	1,333	1,323	56.0	383	1,216	8.4	74.1	507	1,609	11.1
1988	1,333	1,318	60.2	400	1,313	8.2	79.3	527	1,731	10.9
1989	1,333	1,314	62.6	405	1,422	8.5	82.3	533	1,869	11.1
1990	1,333	1,312	64.3	399	1,569	8.5	84.4	524	2,058	11.2
1991	1,333	1,315	63.1	436	1,481	8.1	82.9	573	1,947	10.6
1992	1,333	1,281	66.8	396	1,602	8.1	85.5	506	2,045	10.4
1993	1,333	1,287	66.3	387	1,676	7.9	85.3	496	2,152	10.2
1994	1,333	1,216	72.1	397	1,866	8.2	87.7	481	2,267	9.9
1995	1,333	1,193	80.0	444	2,020	8.1	95.4	526	2,409	9.6
1996	1,333	1,007	99.4	536	2,442	9.3	100.1	533	2,459	9.4
1997	1,124	925	107.7	563	2,764	9.8	99.6	512	2,557	9.1
1998	1,041	835	122.6	618	2,991	10.6	102.4	510	2,494	8.9
1999	975	766	133.4	615	3,499	11.2	102.2	465	2,680	8.6
2000	993	723	147.4	643	4,053	11.7	106.6	457	2,930	8.5

Notes: Summary statistics for firms included in 1986–1996 balanced panel. Available observations from earlier and later years are also incorporated, although substantial sample composition changes limit direct comparison of means and sums. Means and sums for scientists and engineers, sales, and employment variables are calculated over observations with positive R&D. These means and sums will not necessarily add-up with the listed number of observations due to missing values for the particular variable studied. Expenditures are in nominal dollars

OUTUSCOMP, OUTFOREIGN). The relatively complete histories of these core variables facilitate the calculation of R&D stocks and similar lagged investment metrics necessary for accurate productivity analyses. They can thus be paired with annual operating data for econometric exercises that exploit high-frequency and across-firm variations in the data for inference. The operating data can be aggregated to the firm-year level, or the R&D metrics

can be applied to the establishments within a firm with an appropriate clustering of standard errors. One advantage of keeping the data at the plant level is the incorporation of industry and geographic time trends or business cycles that also impact operations.

The second approach concentrates on the smaller number of the large firms that consistently answer the optional questions. These snapshots provide literally hundreds of additional variables on a semi-annual basis. As such, they can be used with the very detailed operating data that are collected at five year intervals. These empirical estimations would focus on lower-frequency variations. In addition to across-firm variation, specifications can also consider within-firm variations by examining R&D efforts in different states or industries. While restricted to the largest firms, exploiting within-firm variation can assist with concerns over firm-level omitted variable biases. Researchers should continue to cross-reference the mandatory variables to understand the firm's activity in years when the optional questions are not administered.

Readers interested in further descriptive statistics and background on the RAD should consult Adams and Peck (1994) and Hall and Long (1999). Kerr and Fu (2006a) also contains additional information exploiting the assignment of firm names to the RAD:

- Steps for cleaning the raw RAD data, aggregating subsidiary records when encountered, reformatting variables, and so on.
- Comparison and reconciliation of RAD firms with Compustat R&D records, with a discussion of differences arising due to the reporting of federally funded R&D.
- Detailed discussion of the impact of corporate restructuring (e.g., mergers, spin-offs) and firm entry/exit for the development of RAD panels.
- Adding-up exercises for the optional break-outs listed above.
- Documentation and verification of major US employers not included in the RAD files.

This article next turns to a description of the NBER patenting database and the matching of this dataset to the RAD.

3 PAT dataset and RAD–PAT matching process

The NBER patent dataset (PAT) was developed by Hall et al. (2001) to facilitate detailed studies of technology growth and diffusion in the US economy. PAT contains over three million individual records for all patents granted by the US Patent and Trademark Office (USPTO) from 1975–2002.⁴ The patent records include company names and identification codes for those assigned to a corporate entity. These records also include a wealth of additional information about the inventions: the application and grant dates, the detailed technology field(s) of the innovation, the inventor name(s), the city and state from which the patent was filed, and citations of prior patents on which the current work builds. Approximately 72% of patents are assigned to firms, with military and government agencies, universities, and unaffiliated applicants accounting for the remainder. Readers should refer to Hall et al. (2001) for comprehensive descriptive statistics on the database.

Firm names are used to match the corporation-affiliated PAT records with the Census Bureau data. However, to minimize disclosure risk and to conserve file size, the Census Bureau does not include firm names in most data files. Instead, firms are identified through alphanumeric ID codes in most Census Bureau datasets. These identifiers facilitate the longitudinal linkages of firm records, the mappings of individual establishments to their

⁴ Patent records are also available prior to 1975 without the inventor name and citation information.

parent organizations, and the cross-merger of datasets within the Census Bureau family. The identifiers are time-invariant except in cases like mergers and acquisitions or the growth from a single-unit to multi-unit firm. The technical report (Kerr and Fu 2006a) discusses the firm name assignments in these cases.

To prepare the RAD for matching with PAT, it is thus necessary to extract firm names from the Census Bureau's Standard Statistical Establishment List (SSEL) Name and Address Files. Derived from tax records, the SSEL and its successor, the Business Register (BR), include the names and addresses of every establishment in the US. Firm names are pulled from the 1987, 1992, and 1997 SSEL single-unit and multi-unit files. The SSEL names are likely more accurate in these economic census years, and these dates approximate the start, middle, and end of the 1986–1996 RAD balanced panel. Firm names listed with the largest establishments are taken for multi-unit firms where differences exist.

This process assigns firm names to every record within the RAD and Longitudinal Research Database. Even before undertaking the PAT matching, incorporating firm names with the RAD affords several quality assurance checks (described above) that are not otherwise possible. The inclusion of firm names across 11 years is also useful for evaluating the dynamic accuracy of the RAD panel and the Census Bureau's longitudinal linkages. Significant name changes further signal corporate restructurings that should be addressed in empirical estimations. To be complete, these mergers and acquisitions are confirmed and supplemented using external vendor data. While the disclosure rules of the Census Bureau do not permit the public release of this information, the quality of RAD estimations is substantially enhanced by incorporating the SSEL firm names. Interested researchers should request both datasets.

The goal of the matching effort is the time-invariant pairings of the Census Bureau's firm ID codes with PAT's assignee codes. The RAD is performed at the firm level, but corporations often file for patents through subsidiaries, legal counsels, and the like. Thus, the mappings are generally multiple PAT assignee codes to a single RAD firm. If a pairing can be made in one year, it can usually be applied forward and back for the full span of the RAD and PAT records. The general challenge of the merger process is not changes in pairings, as both codes are time invariant, but establishing the full set of appropriate mappings when firms develop new assignee codes.

Firm names offer the cleanest and most comprehensive path for making these initial linkages. The names in both datasets are first capitalized. The following standardization procedure is then employed (using a fictitious 'THE O'BRIEN & JOHNSON WIDGET COMPANY USA' as an example):

Step 1: Truncate the initial 'THE ' that starts many company names. The space is included after 'THE ' so that names like 'THERMAL WIDGETS' are not shortened inappropriately in this step.

(O'BRIEN & JOHNSON WIDGET COMPANY USA)

Step 2: Remove any spaces within a name.

(O'BRIEN&JOHNSONWIDGETCOMPANYUSA)

Step 3: Remove the punctuation markers: \ - . & , + “ ” # () / \$.

(O'BRIENJOHNSONWIDGETCOMPANYUSA)

Step 4: Truncate trailing company identifiers (1): AB, AG, BV, CENTER, CO, COMPANY, COMPANIES, CORP, CORPORATION, DIV, GMBH, GROUP, INC,

INCORPORATED, KG, LC, LIMITED, LIMITEDPARTNERSHIP, LLC, LP, LTD, NV, PLC, SA, SARL, SNC, SPA, SRL, TRUST, USA.

(O'BRIENJOHNSONWIDGETCOMPANY⁵)

Step 5: Remove the apostrophe punctuation marker.

(OBRIENJOHNSONWIDGETCOMPANY)

Step 6: Truncate trailing company identifiers (2): CO, COMPANY, CORP, CORPORATION, GROUP, LIMITED, MANUFACTURING, MFG, PTY, and USA. This second truncation accounts for names ending with CO CORP, CO INC, CO LLC, CO LTD, COMPANY CORP, COMPANY INC, PTY LTD, USA INC, and so on.

(OBRIENJOHNSONWIDGET)

A careful review of the primary panel of RAD firms confirms that the above steps, as ordered, do not create multiplicity errors by removing too much information (i.e., making two distinct company names appear the same). Many common leading identifiers, however, should be retained (e.g., 'International', 'United States'). Name-matching algorithms assigning gender or ethnicity to individuals' names often truncate the name length at a specified length (e.g., Kerr 2007). These algorithms are typically less concerned with pairing two names together in a unique mapping, but rather simply the assignment of a population characteristic to them. Experimentation determined this step weakened performance for the unique matching of firm names, however, due to the multiplicity problem.

Automated matching with these standardized names successfully establishes most initial links. The next step is to correct manually simple unmatched cases. Name mismatches are often due to minor complications like typos, abbreviations, and obvious name changes or word re-orderings. This manual alignment also incorporates many subsidiary organizations with a common word stem like 'O'BRIEN & JOHNSON WIDGET R&D LABS'. For the balanced panel, 1,221 of the 1,333 RAD firms are matched to at least one PAT assignee code at this stage (92%).

Even for this matched set, however, some assignee mappings are incomplete due to subsidiaries with distinctly different names. Firm names can also change over time in ways not captured by the three name draws from the 1987–1997 SSELs (e.g., due to an acquisition prior to 1986). Progress towards completing the set of assignee links is first made through external parent-subsidiary links previously established for PAT. Thereafter, extensive searches and business directories further establish the correct linkage for (1) any RAD firm in the balanced panel, (2) any RAD firm among the top 100 R&D performers in a census year but not in the balanced panel, and (3) any PAT assignee code making at least 50 US-filed patents during the 1975–1999 period.

While this manual effort mainly serves to complete the ID-to-assignee mappings for large conglomerates, the searches also located corporate information for 30 unmatched balanced panel firms leading to nine additional matches (1,230 in total). A similar matching rate is achieved for major R&D firms not in the balanced panel. In total, 85% and 70% of US corporate R&D and patenting, respectively, are accounted for by the final pairings.

⁵ Only the last trailer is removed in *Step 4* (e.g., 'USA' in this example). The next-to-last trailer is removed in *Step 6* (e.g., 'COMPANY' in this example). This two-step process, separated by *Step 5*'s removal of the apostrophe punctuation marker, improved overall performance of the algorithm with foreign firm names and name abbreviations.

The accuracy of the name-matching process is also verified through Compustat identifiers previously linked into the Census Bureau data. While these Compustat links can facilitate the merger of external data directly, several limitations for this method exist. Most importantly, privately-held US firms and foreign-listed firms are not included in Compustat; approximately 60% of firm-affiliated patents are linked to US public companies. Second, the Compustat identifiers in both datasets are incomplete and PAT's identifiers are not updated from their initial 1989 draw. Nevertheless, a cross-comparison of the Compustat identifiers does provide confidence that the name-matching approach worked well for the publicly listed US companies in the RAD panel.

Kerr and Fu (2006a) further discusses the matching procedures and SAS coding, documents the manual matches and corrections made, and provides additional quality assurance exercises. The article also details the steps required for aggregating PAT assignee codes to firm-level observations and discusses some further issues with year-to-year mappings. The next section discusses how patents and R&D expenditures are allocated within firms across states and industries.

4 Spatial and industrial allocations

To better study the relationship between firms' innovative efforts and operating performance, firm-level patent counts or R&D stocks are appropriate for some applications, while other empirical exercises require these metrics be allocated spatially or across industries or both simultaneously. This section discusses procedures for these allocations for the RAD-PAT linked dataset.

The spatial allocation of R&D investments by state is fairly straightforward. The detailed RAD breakouts support state-level disaggregations for large firms (e.g., matching Widget's 1995 R&D stock in Massachusetts to Widget's 1995 manufacturing establishments in Massachusetts). Quality assurance exercises confirm these state disaggregations add-up well, with 99% of records having a 5% or less discrepancy. The RAD does not support county or MSA distinctions, although additional Census Bureau records on R&D centers in the Auxiliary Establishment Survey may be of assistance.

The spatial allocation of PAT has greater power. From the USPTO inventor addresses, it is straightforward to develop state and MSA break-outs of each firm's patenting. These patent break-outs can then be linked directly to the Census Bureau data, with MSAs being assigned to plants through their county identifiers. It is also possible to incorporate patents at the establishment level through address matching, comparing the SSEL establishment addresses with the USPTO inventor addresses. This extension facilitates within-MSA spillover analyses. Address matching is much more complex, however, and will be undertaken for individual high-tech industries (e.g., computers, pharmaceuticals) as warranted.

The industrial allocations of R&D and patenting are more complicated. The detailed RAD breakouts disaggregate applied R&D expenditures into approximately forty fields. The RAD produced statistics by product field for odd-numbered years from 1957–1965, 1969–1975, and 1981–1997. Some fields enter and exit the survey (e.g., software is reported separately after 1993); the internal paper discusses these longitudinal changes in greater detail. While these applied R&D fields are not directly linked to the SIC system, the authors developed a crosswalk between the product fields and a mixture of both SIC2 and SIC3 codes that retains as much of the field variation as possible. This concordance is available in Kerr and Fu (2006b). Software is the most challenging field to map due to its application within many fields (e.g., telecommunications equipment); researchers should

carefully consider how it is incorporated. In most empirical applications, the ‘other’ fields should be dropped due to heterogeneity within these miscellaneous categories. In general, researchers should be aware that the industrial disaggregations will be less precise than the spatial mappings.

The industry mappings are also more complicated for patents. The USPTO issues patents by technology categories rather than by industries. Combining the work of Johnson (1999), Silverman (1999), and Kerr (2008), concordances are developed to map the US-PTO classification scheme to the SIC3 framework. While the resulting industry divisions align directly with the Census Bureau structure, patents are assigned probabilistically based upon historical distributions. One promising advantage of patents, however, is that the joint distribution of geography-industry can be studied (e.g., matching Widget’s 1995 computer patenting in Boston to Widget’s 1995 computer manufacturing establishments in Boston).

The within-firm spatial and industrial variation of innovative investments is a promising area for future research, especially when paired with the Census Bureau’s establishment-level operating data in the economic censuses. While the patents and R&D expenditures are not directly linked to operating facilities, the intermediate state and industry disaggregations do provide empirical footholds for many within-firm analyses. Projects can exploit this variation for better quantifying the private and social returns to R&D, for exploring technology diffusion through firm networks, for examining corporate venture capital allocations and parent firm responses, and so on.

5 Conclusions

This article details the construction of a firm-level panel dataset combining the NBER patent dataset with the Census Bureau’s and NSF’s Survey of Industrial R&D. The files are linked through a name-matching algorithm customized for the Census Bureau’s SSEL business registry. This technique can be readily extended to other external datasets researchers wish to link to the Census Bureau data. The developed platform offers an unprecedented view of the R&D-to-patenting innovation process and a close analysis of the strengths and limitations of the R&D survey. Through the Census Bureau’s file structure, R&D can be linked to the operating performances of each firm’s establishments (e.g., Longitudinal Research Database, Longitudinal Business Database), further facilitating innovation-to-productivity studies.

Combining PAT with RAD is important for a complete view of the innovation process. Crudely, R&D expenditures and scientists employed can be thought of as inputs to an innovation production function. Patents, on the other hand, are intermediate metrics of the outputs or effectiveness of these innovative efforts. Together, these two data sources form a more complete view of the technology formation process than they do in isolation. Their combination allows the innovative performance of firms to be compared and contrasted, with one future research output from this project identifying the characteristics of high-productivity research labs in terms of patenting rates. Follow-up research will also carefully quantify the length of the R&D-to-patenting lag and its determinants. The cross-comparison of RAD and PAT is informative to the Census Bureau and NSF, as well, as they work to redesign the RAD’s sampling frame and questionnaires.

With the RAD to PAT linkage established, the Census Bureau’s file structure further facilitates the incorporation of operating data from the Longitudinal Research Database, the Longitudinal Business Database, and other Census Bureau data sources. These operating data allow second-stage analyses of how innovation outputs translate into realized

economic benefits like plant-level output and productivity growth. Specific attention will be given to the types of technologies adopted by plants and the adoption costs associated with these upgrades (e.g., investment expenditures, short-term capacity disruptions, employment upgrading). Working with these establishment data further allows for (1) within-firm comparisons across geographic regions or industries, (2) identifying across-firm spillovers of R&D efforts, and (3) an adding-up exercise to study overall US productivity gains through the intensive changes within companies and the extensive changes of establishment entry and exit.⁶

This project focuses on corporate investments in innovation that can be measured through R&D expenditures and patent grants. Of course, other researchers may be interested in alternative metrics of innovation like copyrights or trademarks. For some industries, especially outside of manufacturing, these metrics may be more appropriate than traditional R&D and patents. The name-matching approach summarized in this article is readily extended to other firm-level datasets as required. In parallel projects, corporate venture capital and corporate restructuring (e.g., mergers and acquisitions, leveraged buyouts) datasets are being linked into the Census Bureau data family through this platform. Interested researchers are welcome to contact the authors about the feasibility of incorporating their own materials in this manner.

In short, the linked dataset described in this article is expected to facilitate comprehensive analyses of the output and productivity gains from R&D investments, using patents as intermediate metrics of the successfulness of R&D efforts. The comparative gains from foreign-sourced or US-outsourced R&D for the operating performances of US establishments can also be quantified. Moreover, the combination of establishment-level operating data and patent citations provides a powerful laboratory for studying technology transfer, knowledge diffusion, and local productivity spillovers (e.g., Ellison et al. 2007). More broadly, the dataset serves as a starting point for macroeconomic research like the impact of US patent regulations on innovation and entrepreneurship.

Acknowledgments The authors thank Jim Davis, Lucia Foster, Adela Luque, and Francisco Moris for assistance on this project. The research in this article was conducted while the authors were Special Sworn Status Researchers of the US Census Bureau at the Boston Census Research Data Center. Research results and conclusions expressed are those of the authors and do not necessarily reflect the views of the Census Bureau. This article has been reviewed to ensure that confidential data are not revealed.

References

- Adams, J., & Jaffe, A. (1994). The span of the effect of R&D in the firm and industry. Center for Economic Studies Working Paper 94-7.
- Adams, J., & Peck, S. (1994). A guide to R&D data at the center for economic studies U.S. bureau of the census. Center for Economic Studies Working Paper 94-9.
- Autor, D., Kerr, W., & Kugler, A. (2007). Do employment protections reduce productivity? Evidence from U.S. states. *Economic Journal*, 117(521), 189–217.
- Cohen, W., Levin, R., & Mowery, D. (1987) Firm size and R&D intensity: A re-examination. *The Journal of Industrial Economics*, 35(4), 543–565.
- Davis, S., Haltiwanger, J., & Schuh, S. (1996). *Job creation and destruction*. Cambridge: MIT Press.
- Ellison, G., Glaeser, E., & Kerr, W. (2007). What causes industry agglomeration? Evidence from coagglomeration patterns. NBER Working Paper 13068.

⁶ Autor et al. (2007) and Kerr and Nanda (2007) offer examples with employment protection and banking regulations, respectively.

- Hall, B., Jaffe, A., & Trajtenberg, M. (2001). The NBER patent citation data file: Lessons, insights and methodological tools. NBER Working Paper 8498.
- Hall, B., & Long, W. (1999). Differences in reported R&D data on the NSF/Census RD-1 form and the SEC 10-K form: A micro-data investigation. Working Paper.
- Jaffe, A., Trajtenberg, M., & Fogarty, M. (2000). Knowledge spillovers and patent citations: Evidence from a survey of inventors. *The American Economic Review*, 90(2), 215–218.
- Jarmin, R., & Miranda, J. (2002). The longitudinal business database. Center for Economic Studies Working Paper 02-17.
- Johnson, D. (1999). 150 Years of American invention: Methodology and a first geographic application. Wellesley College Economics Working Paper 99–01. Data currently reside at <http://faculty1.coloradocollege.edu/~djohnson/uships.html>.
- Kerr, W. (2007). The ethnic composition of US inventors. Harvard Business School Working Paper 08-006.
- Kerr, W. (2008). Ethnic scientific communities and international technology diffusion. *Review of Economics and Statistics* (forthcoming).
- Kerr, W., & Fu, S. (2006a). The RAD-patent-LRD mapping project. Census Bureau Technical Report.
- Kerr, W., & Fu, S. (2006b). The industry R&D survey—patent database link project. Harvard Business School Working Paper 07-031.
- Kerr, W., & Nanda, R. (2007). Banking deregulation, financing constraints and entrepreneurship. Harvard Business School Working Paper 07-033.
- Levin, R., Cohen, W., & Mowery, D. (1985). R&D appropriability, opportunity, and market structure: New evidence on some Schumpeterian hypotheses. *The American Economic Review*, 75(2), 20–24.
- Silverman, B. (1999). Technological resources and the direction of corporate diversification: Toward an integration of the resource-based view and transaction cost economics. *Management Science*, 45(8), 1109–1124.