70 Years of Patents Matched to Firms:
Methodology and Insights About Firm Heterogeneity

Arnaud Dyèvre Oliver Seager

June 2023 – Version 0.1
(Latest version – Insert link)
Please to not share or cite

ABSTRACT. This paper presents a novel dataset that matches patent data from the US Patent Office with Compustat data over seven decades. Building upon previous efforts, we improve existing panels of Compustat firms and patents in three significant ways. Firstly, we extend the temporal coverage, doubling the length of the most comprehensive panel. Secondly, we meticulously investigate historical changes in corporate structure, including mergers, spin-offs, and relistings, ensuring accurate tracking of patent ownership over time. Thirdly, we rectify false positives in existing datasets through a combination of manual and automated matching techniques, resulting in a final panel comprising 9,708 unique firms matched to 3,448,337 USPTO patents, the most comprehensive of its kind. We make the dataset and replication code publicly available for researchers to use and improve upon. We then demonstrate the research potential of the dataset by using its two key features: its time coverage and the dynamic matching of patents to firms over time. We document facts about the quality of innovation performed by firms of different sizes and the effects of mergers on the innovation trajectories of firms.

Keywords: Patent data, Innovation, Firm heterogeneity, R&D
JEL code: O3

Dyèvre: LSE, Department of Economics. Seager: LSE, Department of Management
For their valuable suggestions and comments, we are grateful to Abhijit Tagade, [To be extended].
This project benefited from the financial support of STICERD (grants #107728 and #108954), which is gratefully acknowledged.

1. Introduction

Comprehensive data on the innovative outputs of firms is central to the study of corporate innovation dynamics, long-run productivity growth and technological change. One of the most commonly used types of data in growth and innovation studies has been panels of firms with balance sheet data matched to patents. They have been used to study, among others, technology spillovers between firms (Bloom et al., 2013), the balance between incremental improvements of a firm's existing products and creating brand new ones (Akcigit and Kerr, 2018), the use of science in corporate innovation (Arora et al., 2021a) and the impact of technology standardization on firm performance (Bergeaud et al., 2022). However, long-run analyses of firms' patenting trajectories have been held back by the lack of historical data combining both patent information and observables on firms (such as employment, sales, capital expenditures and R&D spending). Patent datasets such as those provided by the US and the European patent office contain the names of patent assignees–the entities who won the monopoly rights granted by patents–but assignee names are reported as string variables, as they are entered by the assignee filing the patent. This creates issues for linking patent filings with firm performance as the name of the firm in the patent data ('International Business Machines' for instance) may not match directly to the name reported in the firm balance sheet dataset ('IBM').

In this paper, we present the methodology we used to create the most comprehensive dataset of publicly listed firms matched to patents. Our goal is to facilitate the use of this data by researchers interested in innovation and growth. The dataset covers all publicly listed firms whose balance sheet information is available in Compustat North America, and who have been granted patents by the US Patent and Trademark Office (USPTO), from 1950 to 2020. The final sample of firms is an unbalanced panel dataset of 9,233 unique firm identifiers ('gvkey'), which account for 82,314 firm$\times$year observations matched to 3,188,444 patents. Because some patents are filed by private firms or lone inventors before being acquired by publicly listed firms later on, the total number of unique patents that get matched to a publicly listed firm at some point is 3,448,337, and the number of firms to which a patent is at some point assigned is 9,708. If we further incorporate patents granted by the USPTO during 1926-1949 in our sample, these figures increase to 3,631,589 patents and 9,821 firms, respectively. Importantly, our dataset re-assigns patents to firms when a merger, an acquisition, a name change or a relisting occurs. This allows one to correctly track the patent stocks of firms even if patents change hands, are aggregated up to a larger entity or are split across subsidiaries. To build this dataset, we combine data from X sources: (i) the patent data from 1976 to 2020 comes from PatentsView, (ii) patent data from 1950 to 1975 come from Fleming et al. (2019), (iii) firm balance sheet data come from Compustat North America, (iv) corporate restructuring information comes from SDC platinum and online searches, finally (v) we also track companies' name changes through WRDS' CRSP data. This dataset is freely available for researchers to use and to improve upon. Data, annotated Stata and Python code, and methodology can be accessed here: `github.com/arnauddyevre/compustat-patents`. We will update the data annually

and we welcome suggestions from researchers to improve it.

Our dataset presents two distinctive features. Firstly, it tracks corporate ownership changes–through M&A, spinoffs and relistings–from 1950 to 2020, thus allowing one to appropriately measure the patent stock of a firm who has gone through any form of corporate restructuring. Our contribution is to combine a database of M&As (SDC Platinum) with comprehensive manual searches and refinements for historical changes in corporate structure which have not been used before and that we make available as part of this project. This allows us to list X such changes in corporate structure over seven decades. Our manual incorporation of historical corporate events constitute the main contribution of our data (beyond its unique time coverage). A second distinctive feature of our data is that it contains citation data from all patents granted between 1950 to 2020, allowing us to leverage patent citations to study spillovers over the long run and computing citation-based measures of quality.

Our dataset extends the coverage of previous firm-patent matching efforts. Most notably that of Arora et al. (2021b) and Hall et al. (2001). We are able to match 139% more patents to firms and include 109% more firms in our panel as Arora et al. (2021b) do. Our panel doubles the time coverage of previous efforts. It also identifies inadequate matches between patent assignee names (in the patent data) and firm names (in Compustat), either false positives or false negatives. Within the time period covered by previous dataset, we identify X% false negatives (Compustat firms not matched to patent assignees while they should) and X% false positives (firm names in Compustat matched to assignees in the patent data while they should not).

In the second part of the paper, we illustrate the research potential of this data by revisiting stylised facts about firm heterogeneity and innovation (Cohen, 2010) and evaluating their stability over the full duration of our dataset. We also leverage the dynamic matching of patents in our data to document the impact of mergers and acquisition on firms' innovation trajectories.

Related literature. This paper relates to two separate strands of literature, one methodological, the other analytical. On the methodological front, this paper relates to the previous efforts of other researchers who have build panels of firms matched to patents.

We are not the first to develop a dataset of Compustat firms dynamically matched to patents. Most notably, Ashish Arora, Sharon Belenzon and Lia Sheer have made available a database of Compustat firms dynamically matched to patents covering the period 1980-2015. Their work, in turn, builds upon the previous efforts of Bessen (2009) who first included some dynamic reassignment of patents in the original (and widely used) NBER patent data (Hall et al., 2001). In section 4, we lay out how our dataset compares and improves upon these important previous efforts.

Regarding the relationship between firm size and innovation, a large and old literature has documented a set of stylised facts about how innovation scale up with firm size. This literature is reviewed in Cohen (2010) and in section II of Akcigit and Kerr (2018).

2. Data sources

To create our panel, we rely on 8 data sources.

(1) **Compustat** provides balance-sheet data for publicly traded firms in North America from 1950 to 2020.

(2) **PatentsView** provides data on digitised patent documents from 1976 to 2020. PatentsView apply extensive homogenisation algorithms to assignee names, consolidating under a single string the various names with which a corporate entity files patents. We use PatentsView as our primary source of patent data for 1976 to 2020.

(3) **Fleming et al. (2019)** collate data from optical character recognition (OCR) scans of patent documents from 1926 to 1975, and digitised patent records from 1976 to 2017. Their data provides our source of patent data for 1926 to 1975. Fleming et al. (2019) also harmonise assignee names across the 1926 to 2017 period. Leveraging this harmonisation and overlapping data with PatentsView from 1976 to 2017, we extend PatentsView's assignee name harmonisation to patents from 1926 to 1975.

(4) **The Centre for Research and Security Prices (CRSP)** document all trading names of publicly traded firms in North America for 1926 to 2020. CRSP also provide a linkage table between their data and Compustat for 1950 to 2020. Compustat only captures contemporary trading names of firms. Through its tracking of name changes, CRSP allows us match patents to firms through their previous trading names. Additionally, we use the CRSP-to-Compustat linkage table to track M&A and corporate restructuring activity for the 1950 to 2020 period.

(5) **SDC Platinum** offers worldwide data on mergers and acquisitions becoming effective between 1985 and 2020. We utilise SDC Platinum to track M&A activity, where both acquirer and target are publicly traded, from 1985 to 2020.

(6) **Arora et al. (2021b)** include in their data the names of privately-held subsidiaries of publicly traded firms from 1980 to 2015. They also provide a linkage table between their data and Compustat for the same period. We link patents to ultimate-owner firms through subsidiaries using the subsidiary ownership data Arora et al. (2021b) provide. Furthermore,

we utilise the linkage table between their data and Compustat to document M&A and corporate restructuring events from 1980 to 2015.

(7) **WRDS Company Subsidiary Data** provides subsidiary names listed in SEC 10-K filings, submitted annually by all publicly traded firms in the United States, from 1993 to 2019. These data augment the catalogue of privately-held subsidiaries through which we match patents to publicly traded firms. Additionally, we use these data to identify M&A activity when a firm exits Compustat and it, or one of its subsidiaries, first appears as a subsidiary of another Compustat firm in the same or following year.

(8) **Lev and Mandelker (1972)** include in their data M&A events, where publicly traded firms are acquirers, from 1952 to 1963. To our knowledge, existing databases and online sources documenting acquisitions of privately-held subsidiaries are scarce in their coverage of the 1950s and early 1960s. Thus, Lev and Mandelker (1972) provide a critical expansion to our catalogue of subsidiaries.

(9) **Manual searches on mergers and acquisitions** We add X manually curated "corporate reassignment events" such as mergers, acquisitions, re-listings and spinoffs to the database. This is essential to cover the earlier years of our data, which are not covered by SDC Platinum (starting in 1985). This manually curated list of events is available on the project page.

Compustat notwithstanding, we thoroughly review each data source through manual inspection and online searches. This review allows us to make several manual additions to our data, including patent assignee name homogenisations, M&A events, and corporate restructuring events.

3. Data construction

In our matching of patents to balance sheet data, we map two distinct sets of patents to each observation in the Compustat Fundamentals Annual panel of firms. The first set is our *static match*, capturing the link between assignees and their patents at the time of filing the patent. The second set is our *dynamic match*, capturing patent stock: this comprises all patents generated by the firm in previous years.[1]

If corporate structure were immutable over the life-cycle of a firm, each dynamic match would be the union of all past static matches. Economic reality, however, presents a more complex picture of firm entry and exit: a firm can undergo a merger, acquire subsidiaries or itself be acquired, spin-off subsidiaries or be born of a spin-off, or can divide into many firms at the behest of regulatory authorities. Adding further complexity to dynamic matching, the GVKEY identifying a firm

[1]We describe a firm as 'generating a patent' with deliberate ambiguity here. Our data facilitates the matching of patents to firm-years either by either (i) the date of application submission for later-granted patents or (ii) the date of patent granting. Each researcher can decide which is more appropriate for their purposes.

in Compustat often changes upon alterations to the architecture of a firm's equity, even if these alterations do not impact its stock of intellectual capital. This can occur where a firm is taken private following a leveraged buyout and later relisted, where a firm is de-listed from one exchange and relisted on another, or where a firm becomes a subsidiary of a holding company solely for accounting purposes. Compustat does not provide linkages between the relevant GVKEYs in such cases. Therefore, whilst built on top of our static match, our dynamic match requires additional methodological steps required to track a firm's patent base over time in light of changes to ownership structure. As such, we deal with the construction of each match separately.

Before explicitly detailing the construction of our static and dynamic matches, we elaborate on two key components of our data construction: our name standardisation algorithm, and our treatment of patent data.

3.1. **Name Standardisation.** USPTO patent documents identify the firm to whom a patent is assigned by name only. As such, a firm's name provides the sole channel through which we automatically match patents to Compustat GVKEYs.

This necessitates the standardisation of firm names, which may appear under different strings in patent data and balance-sheet data. For example, aerospace manufacturer Boeing is assigned patents as 'THE BOEING COMPANY', and is listed in Compustat as 'BOEING CO'. To reconcile such differences, we apply a six-step standardisation algorithm to our universe of firm names. This proceeds as follows:

(1) Extraneous whitespace is stripped from all strings. This includes removing leading and trailing blanks, and replacing multiple consecutive whitespace characters with a single whitespace character.

(2) Non-alphanumeric characters (e.g. commas, periods) and nonstandard alphanumeric characters (e.g. Ä, É) are replaced with a corresponding alphanumeric character, or removed altogether. Examples of this are given in Table 1.

(3) Acronyms are 'condensed', with all whitespace between standalone non-whitespace characters removed. Examples of this are given in Table 2.

(4) Generic corporate suffixes and 2-letter state codes are removed. Examples of this are given in Table 3.

(5) Commonly abbreviated terms are replaced with a standard abbreviation thereof. Examples of this are given in Table 4

(6) Any remaining whitespace is removed.

TABLE 1. Second step of our name standardisation algorithm: replacing or removing non-alphanumeric and nonstandard alphanumeric characters.

Name Before Alphanumeric Standardisation	Name After Alphanumeric Standardisation
WITTE + SUTOR GMBH	WITTE AND SUTOR GMBH
SWS ENGINEERING S.P.A.	SWS ENGINEERING S P A
FORTUMO OÜ	FORTUMO OU
AERO-DRI CORPORATION	AERO DRI CORPORATION

TABLE 2. Third step of our name standardisation algorithm: condensing acronyms

Name Before Acronym Condensation	Name After Acronym Condensation
SWS ENGINEERING S P A	SWS ENGINEERING SPA
SCIENZ GROUP L L C	SCIENZ GROUP LLC
F M HOWELL AND COMPANY	FM HOWELL AND COMPANY
SIGNTECH U S A LTD	SIGNTECH USA LTD

TABLE 3. Fourth step of our name standardisation algorithm: removing generic corporate suffixes and 2-letter state codes

Name Before Generic Term Removal	Name After Generic Term Removal
SWS ENGINEERING SPA	SWS ENGINEERING
SCIENZ GROUP LLC	SCIENZ GROUP
FM HOWELL AND COMPANY	FM HOWELL
BROCKWAY INC NY	BROCKWAY

TABLE 4. Fifth step of our name standardisation algorithm: mapping commonly abbreviated terms to a standardised abbreviation.

Name Before Standardised Abbreviation	Name After Standardised Abbreviation
SWS ENGINEERING	SWS ENG
SCIENZ GROUP	SCIENZ GP
PRESSER INTERNATIONAL	PRESSER INT
COACH MASTER INTL	COACH MASTER INT

In sum, our standardisation algorithm alters 98.9% of firm names in our patent data, and 99.7% of names in our balance-sheet data.

3.2. **Further Treatment of Patent Data.** Beyond algorithmic standardisation of firm names, we take additional steps to ensure that the firm names associated with patents are both accurate and easily matched to balance-sheet data.

The OCR techniques used by Fleming et al. (2019) to parse firm names from 1926-1975 patent documents are subject to certain limitations. Such a procedure must not only translate images of a patent document to text, but also determine which portion of the text pertains to which attribute of the patent. Whilst generally Fleming et al. (2019) correctly identify firm names, many names are erroneously given of a form akin to 'Assignors to Reliance Electric and Engineering of Ohio Application March 22 1947 Serial No. 736532' instead of 'Reliance Electric and Engineering'. To rectify this issue, we identify firm names that begin with 'ASSIGN' or a similar substring.[2] For each of these, we find the longest substring appearing directly after the substring 'TO' that matches a firm name elsewhere in the dataset. We manually review each of these matches, and replace the firm names associated with relevant patents accordingly.

Both PatentsView and Fleming et al. (2019) develop their own assignee name harmonisation algorithms to clean their patent data. These algorithms consolidate various versions of a firm name under a single, canonical string. Since we draw on both datasets, however, we have 1926-1975 patents associated with Fleming et al. (2019)'s harmonised firm names, and 1976-2020 patents associated with PatentsView's harmonised firm names. Any difference in harmonisation algorithms risks a discrepancy between the canonical string associated with a firm's patents in each dataset, creating a discontinuity in patenting activity between 1975 and 1976 when matching to balance-sheet data. To remedy this, we leverage the 1976-2017 coverage common to both datasets: using patents from this period, we create a mapping from harmonised names in the Fleming et al. (2019) data to harmonised names in the PatentsView data. Where possible, we then use this mapping to associate patents from the 1926-1975 period with the relevant firm name from the PatentsView harmonisation. An example of this is given in Table 5.

Building on the PatentsView harmonisation of firm names, we subject all firm names associated with fifty or more patents to further review. Within this restricted sample of names, we manually review all pairs for which one name is a substring of the other (e.g. 'Philips' as a substring of 'Koninklijke Philips') and consolidate relevant patents under a single firm name accordingly. Furthermore, for the 250 firm names associated with the most patents, we conduct data review and online searches to find alternative names associated with each firm.

[2]Specifically, any firm name whose first six characters have a maximum Levenshtein distance of one from 'ASSIGN'.

TABLE 5. Extending harmonised names from PatentsView data to 1926-1975 patents

Patent Number	Year Granted	Standardised Name: Fleming et al. (2019)	Standardised Name: PatentsView	Standardised Name: Final
3828671	1974	MEDALISTIND		MEDALIST
3849801	1974	MEDALISTIND		MEDALIST
3909847	1975	MEDALISTIND		MEDALIST
4062157	1977	MEDALISTIND	MEDALIST	MEDALIST
4056137	1977	MEDALISTIND	MEDALIST	MEDALIST
4084504	1978	MEDALISTIND	MEDALIST	MEDALIST
4354768	1982	MEDALISTIND	MEDALIST	MEDALIST
4771651	1986	MEDALISTIND	MEDALIST	MEDALIST

Our collated patent data, spanning the 1926-2020 period, comprises 8,651,808 patents associated with 633,530 standardised firm names. The data construction process is summarised in Figure 1

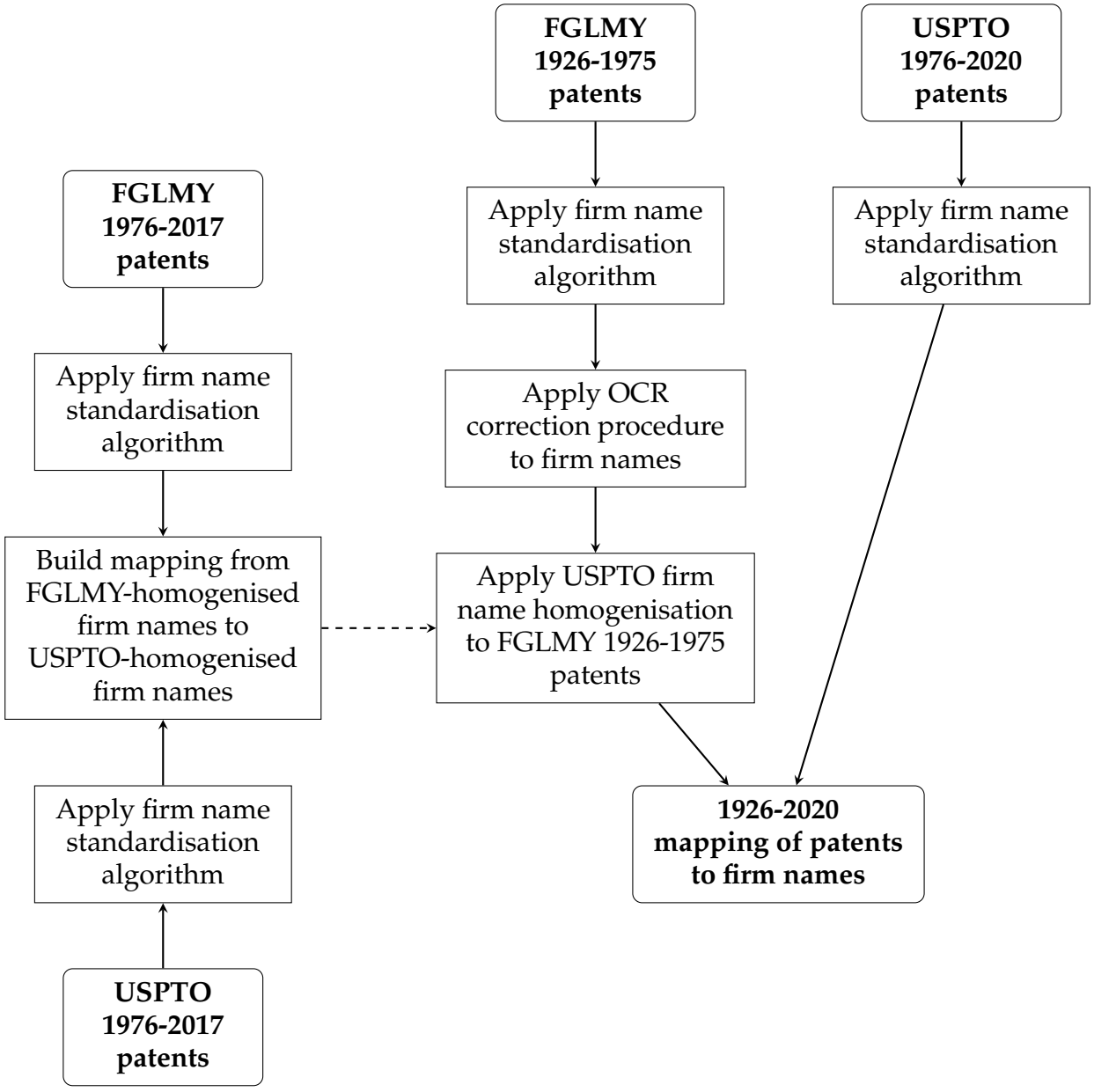
3.3. **Static Match.** Our *static match* maps each GVKEY-year observation in the Compustat panel to the set of patents generated by the firm in the given year. This necessitates constructing an intermediary set of associated firm names for each GVKEY-year, wherein each name should be associated with at most one GVKEY in a given year.

To construct these intermediary sets of names, we firstly concern ourselves with all historical trading names of a firm.[3] Compustat lists only the most recent trading name associated with a given GVKEY. However, our patent data often give a patent's assignee as the name under which the firm traded at the time of application. Therefore, we follow Arora et al. (2021b) in sourcing firm names using the CRSP Daily Stock file and CRSP-Compustat Linking Tables. Using CRSP data, we are able to track all historical names for each GVKEY and the period for which each name was used. 38% of all GVKEYs in our sample have at least two distinct trading names during their tenure in Compustat. Examples of this are given in Table 6.

Secondly, we follow Bessen (2009) in attributing patents to the highest possible level in a corporate structure. As such, we compile data on privately-held subsidiaries for each GVKEY. Our foremost source for this data is WRDS, who match GVKEYs to subsidiary names listed in SEC 10-K filings for 1993 to 2019. For earlier in our sample period, we utilise subsidiary data provided by Arora et al. (2021b) for 1980 to 2015, data on acquisitions of private subsidiaries for 1952 to 1963 provided by Lev and Mandelker (1972), and extensive manual investigation into purpose-created

[3]We use the term 'trading name' here to refer to the name under which a firm has been traded on a stock exchange, which is not to be confused with the name under which a firm conducts its client-facing operations.

FIGURE 1. Patent data construction.



subsidiaries and acquisitions of private firms for the 1950 to 1989 period. In total, this procedure maps 592,635 subsidiaries to 10,245 GVKEYs. Such a large number of subsidiaries creates concerns regarding the potential for erroneous matches. Therefore, we apply several restrictions to our catalogue of subsidiaries, *viz.* (i) dropping any subsidiary with a standardised name that maps to multiple GVKEYs in a single year, (ii) dropping any subsidiary whose standardised name

TABLE 6. Historical trading names of firms

GVKEY	Firm Name	Applicable Period	Source
063083	Cardiovascular Dynamics Inc	1996-1999	CRSP
063083	Radiance Medical Systems Inc	1999-2002	CRSP
063083	Endologix Inc	2002-2019	Compustat
286433	Oasmia Pharmaceutical A B	2013-2019	CRSP
286433	Vivesto AB	2019-2020	Compustat

is identical to a standardised trading name of any GVKEY in our data, (iii) dropping any subsidiary whose standardised name is four characters or fewer, (iv) following Arora et al. (2021b), keeping only subsidiaries of GVKEYs for whom at least one name from Compustat or CRSP can be matched to patents, (v) removing subsidiaries with standardised names that are common corporate or branding terms, such as 'Production' and 'Specialties', and (vi) dropping any subsidiary that cannot be matched to a patent within a decade of its tenure as a subsidiary. These restrictions constitute a substantial reduction to our catalogue, leaving 9,263 subsidiaries that map to 2,239 GVKEYs.

Many publicly traded firms are themselves subsidiaries of other publicly traded firms. As such, we draw on a variety of sources to track chains of ownership between GVKEYs that appear in Compustat. This further enables us to match patents to the listed firm occupying the highest level within a corporate structure. Whilst we leverage this data in our static match, for expositional clarity its construction is outlined in subsection 3.4.

All names matched to GVKEY-years are subject to the name standardisation algorithm detailed in subsection 3.1. In sum, our static match links 3,188,444 patents to 82,314 GVKEY-years, and to 9,233 GVKEYs.

3.4. **Dynamic Match.** Our *dynamic match* maps each GVKEY-year observation to the set of patents generated by the firm in preceding years. Here, complexity arises due to the mutable nature of corporate structures. Events in the life-cycle of a firm sometimes necessitate patents associated with one GVKEY in our static match being later associated with another in terms of historical patenting activity. We term any event necessitating such patent reassignment an 'effective acquisition' between GVKEYs, with this term encompassing the following cases: for acquisitions, we say the acquiror's GVKEY 'effectively acquires' that of the acquiree; in the case of a merger, we say the surviving GVKEY 'effectively acquires' the GVKEY that exits Compustat; if a firm delists and relists its stock, we say that the firm's new GVKEY 'effectively acquires' the old one in the year of relisting; when a publicly-traded parent firm lists a subsidiary, we say the parent's GVKEY 'effectively acquires' that of the subsidiary; when a publicly-traded parent firm spins off

a publicly-traded subsidiary, we say the subsidiary's GVKEY 'effectively acquires' itself from its former parent. We classify each effective acquisition in our data according to transaction type, permitting the separation of true M&A activity from cases of nominal corporate restructuring. The sources we review to compile data on effective acquisitions are various.

For M&A activity specifically, we utilise SDC Platinum, which offers data on realised mergers and acquisitions for 1985 to 2020. We match SDC Platinum's records on acquirors and acquirees to Compustat GVKEYs using 6-character CUSIPs. Because a 6-character CUSIP can be reassigned to different firms over time, each match requires further verification. Therefore, we apply our standardisation algorithm to the names of acquirors and acquirees in SDC Platinum, and manually review all cases where these do not exactly match their counterparts in Compustat. This procedure provides information on 186 effective acquisitions. For effective acquisitions more generally, we manually review each instance in which a 6-character CUSIP is shared by multiple GVKEYs. For example, this enables us to link the relisting of Peoples Jewellers in 1981 (GVKEY 014314) to their previous listing from 1960 to 1975 (GVKEY 008472). In total, manual review of 6-character CUSIPs provides an additional 86 effective acquisitions.

As a further source of effective acquisitions, we appeal to the CRSP-to-Compustat crosswalk. The CRSP Daily Stock File uses two identifiers for each observation: PERMCO, which is fixed at the level of a publicly-traded security, and PERMNO, which is fixed at the firm level. This dual identification allows researchers to track different securities issued by the same firm over time. Using CRSP's PERMNO-GVKEY crosswalk, we identify cases in which a PERMNO undergoes a change in the GVKEY to which it maps. We manually review each case to determine which are indicative of an effective acquisition. For example, this procedure identifies the 2001 merger of offshore drilling firms Santa Fe International and Global Marine: in 2001, PERMNO 85080 (Santa Fe International) ceases to map to GVKEY 064876 (Santa Fe International) and newly maps to GVKEY 005187 (GlobalSantaFe, previously Global Marine). In sum, this procedure yields a further 215 effective acquisitions. We apply the same procedure to data provided by Arora et al. (2021b), who match patents to firms via their own firm identifier, PERMNO_ADJ, providing a similar crosswalk between this and GVKEY. This yields an additional 26 effective acquisitions.

We also identify effective acquisitions through duplicates of trading names created by our name standardisation algorithm. 16.1% of standardised trading names map to more than one GVKEY, compared with 0.8% of non-standardised trading names. Restricting our attention to standardised names which can be matched to patents, we manually review each instance of duplication for indications of an effective acquisition. For example, we identify the 1967 merger of Hajoca (then GVKEY 005428) and Gable Industries (GVKEY 005429) since both GVKEYs carry the standardised name "HAJOCA" for part of their tenure in Compustat. In total, this procedure facilitates identification of a further 245 effective acquisitions.

Leveraging subsidiary data compiled by WRDS from SEC 10-K Exhibits 21, we identify further effective acquisitions in instances where a formerly listed firm becomes the subsidiary of a currently listed firm. We manually review all occasions on which a standardised name of a GVKEY that exits Compustat first appears as a subsidiary of another GVKEY in the year of, or the year following, the former's exit. For example, De Soto Inc. (GVKEY 003888) exits Compustat in 1995. In 1996, De Soto Inc. is listed as a subsidiary of Keystone Consolidated Industries (GVKEY 006424), indicative of Keystone's 1996 acquisition of De Soto. Extending this process, we flag instances in which a subsidiary of a GVKEY that exits Compustat appears as a subsidiary of another GVKEY in the same or following year, subjecting these events to the same manual review. This procedure provides an additional 227 effective acquisitions for our data.

Finally, we review informal online compendia of high-value and industrially significant M&A activity to identify further effective acquisitions. This review focuses especially on the 1950 to 1989 period that has little or no coverage in SDC Platinum. We do, however, check that our other sources cover particularly high-value M&A activity for the 1990 to 2020 period of our sample. This general research provides another 625 effective acquisitions. In sum, our sources identify 993 mergers and acquisitions, 327 relistings, 224 subsidiary listings, and 66 subsidiary spin-offs.

Collating these effective acquisitions, we are able to construct chains of ownership between GVKEYs for our entire sample period. Using this data, we finalise our static match, attributing patents to the highest possible GVKEY in a corporate structure at the time of patenting. In our data, we term this the GVKEYUO (or ultimate owner GVKEY).

For our dynamic match, the GVKEYUO would be a sufficient variable with which to reassign patents if only entire corporate structures could be acquirors or acquirees. However, a subsidiary can be uncoupled from its parent firm through a spin-off or sale to another firm. Since our static match attributes patents to the highest GVKEY in a corporate structure at the time of patenting, we require a means of transferring to any future parent the ownership of patents assigned to a subsidiary. For example, USPTO patent 4596462 (granted to Beckman Instruments, Inc. in 1986) bears the GVKEYUO corresponding to SmithKline Beckman, the firm's parent at the time of application and granting, in our static match. However, Beckman Instruments was spun off as Beckman Coulter in 1989 and acquired by Danaher in 2011. We cannot correctly reassign this patent using its GVKEYUO in the static match, as this corresponds to the firm that, after divesting itself of Beckman Instruments in 1989 and merging with The Beecham Group, continued life as SmithKline Beecham. In order to correctly reassign this patent to Beckman Coulter in 1989 and Danaher in 2011, we need a second firm identifier with which to associate the patent. To this end, we create a GVKEY-like identifier which we term the GVKEYFR (or GVKEY for reassignment). We also create a GVKEYFR-to-GVKEY mapping covering our entire sample period, permitting the matching of GVKEY-years to patenting histories.

FIGURE 2. Corporate structure data construction.

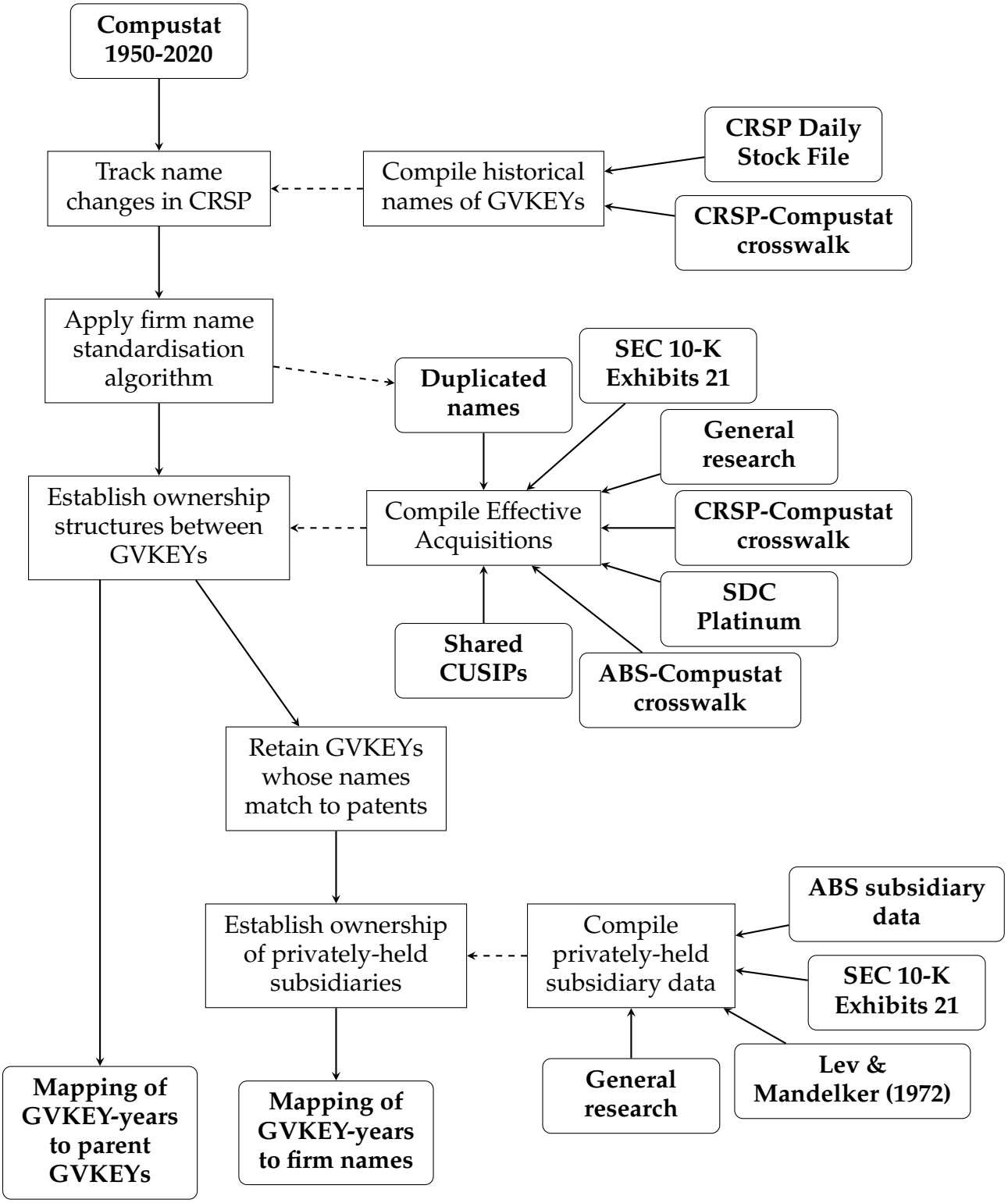


FIGURE 3. Static Match.

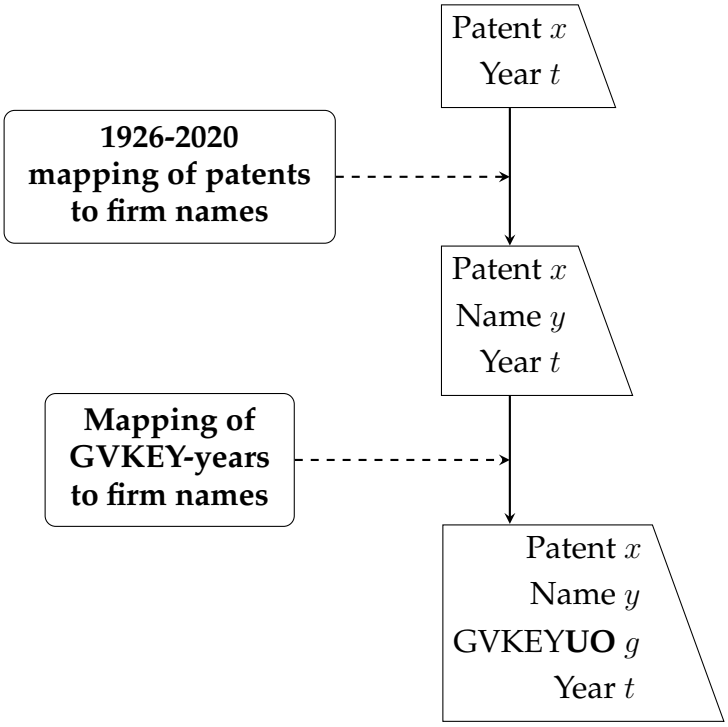


TABLE 7. Example of patent-level data, by application year

Patent Number	Application Year	GVKEYFR	GVKEYUO
2278211	1940	005047	n/a
4356681	1980	009626	008543
4715664	1986	019661	019661
6363733	2000	201140	201140

In sum, our dynamic match tracks the ownership of 3,644,430 patents among 9,821 GVKEYs. An example of our patent-level data is given in Table 7, and an example of our GVKEYFR-to-GVKEY mapping data is given in Table 8.

4. COMPARISON WITH EXISTING DATASETS

4.1. **Shares of patents in DS.**

4.2. **Comparison with existing datasets.**

FIGURE 4. Dynamic Match.

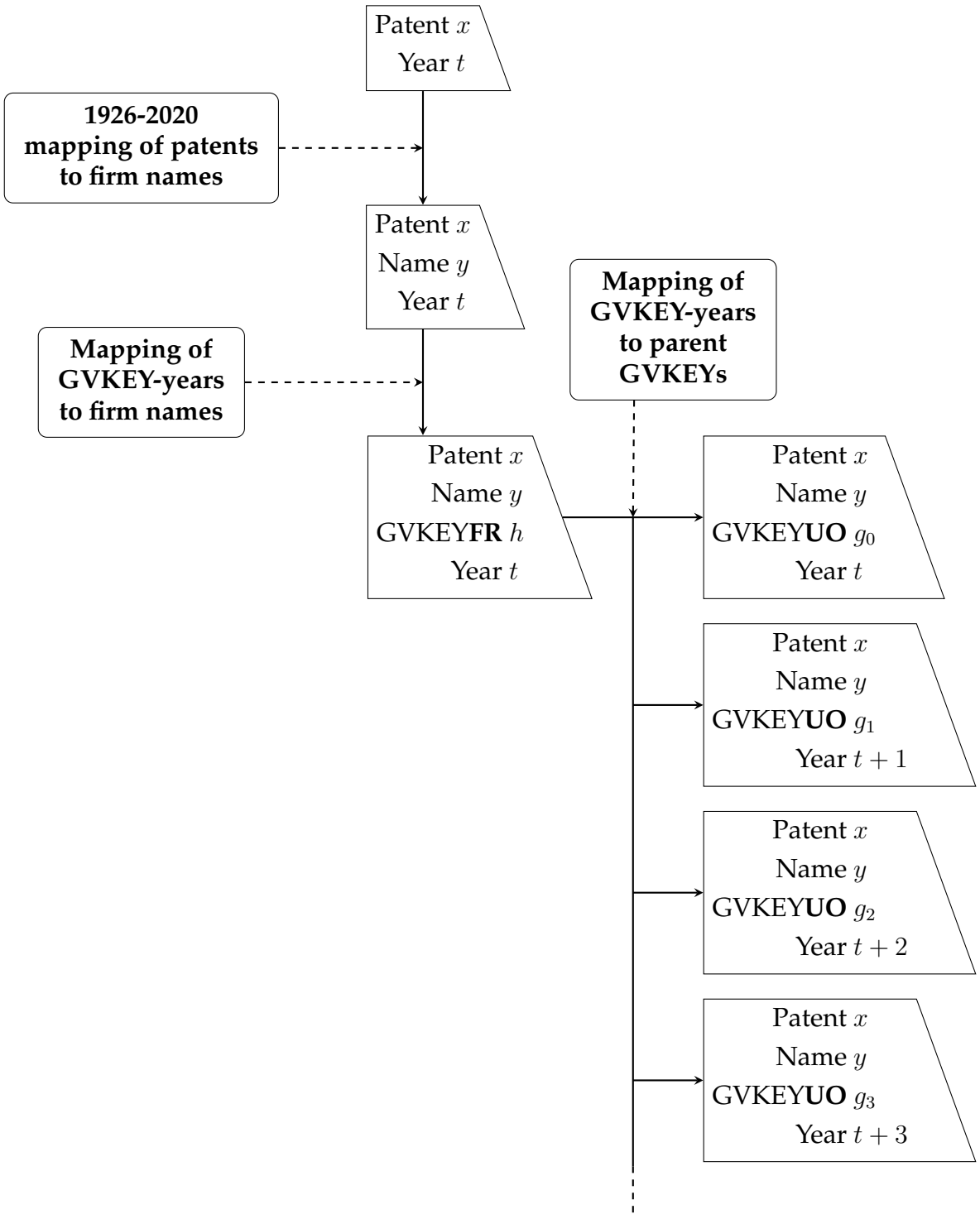


TABLE 8. Example of GVKEYFR-to-GVKEY mapping data

GVKEYFR	GVKEY	First Link Year	Last Link Year
005047	005047	1950	2020
009626	009626	1966	1977
009626	008543	1978	1985
009626	014448	1986	1990
019661	019661	1981	2020
201140	201140	1995	2004
201140	241637	2005	2020

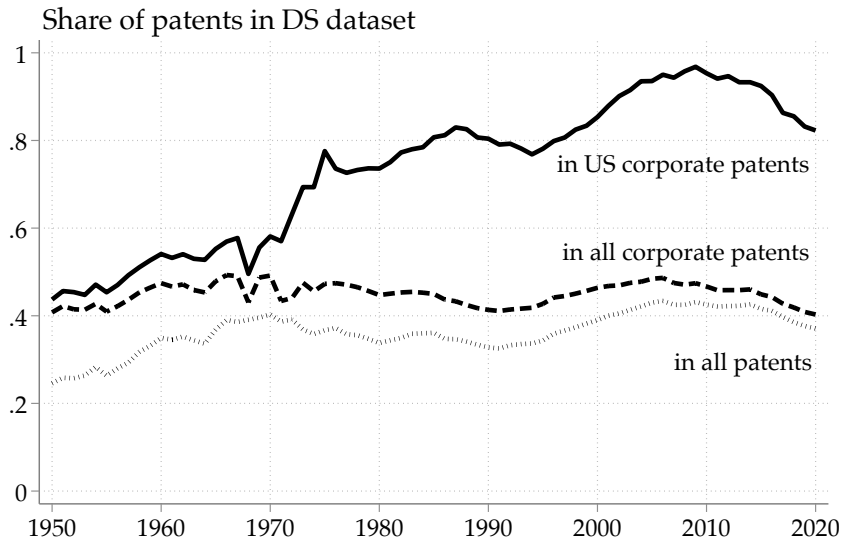


FIGURE 5. Patent coverage of the dataset

Notes: This figure

5. INSIGHTS ABOUT INEQUALITY BETWEEN FIRMS

In this section, we showcase the research potential of our dataset by revisiting two long-standing questions in the economics of innovation: how does innovation scale with firm size? We also present evidence of the change in the nature of innovation and firm performance after an acquisition, in the consolidated firm.

5.1. **Inequality in innovation v. inequality in sales.**

5.2. **Innovation and firm size.**

5.3. **Innovation dynamics after an M&A.** Event studies:

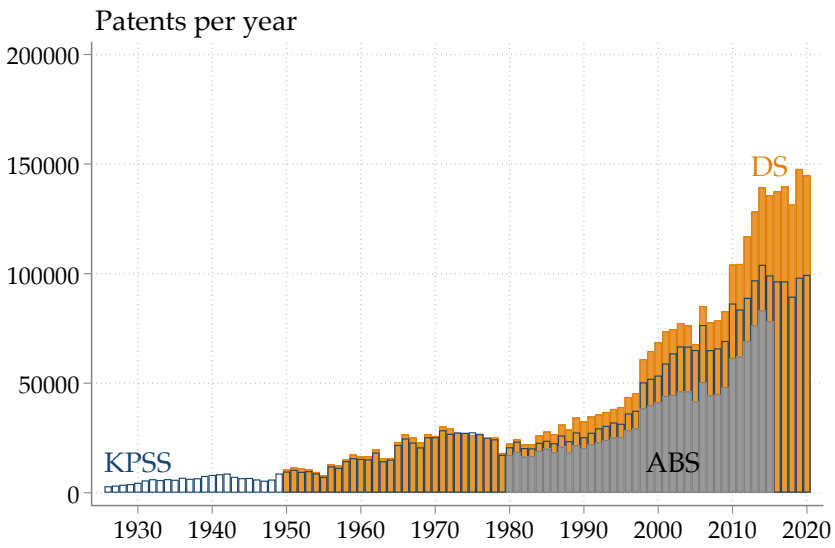


FIGURE 6. Comparison

Notes: This figure

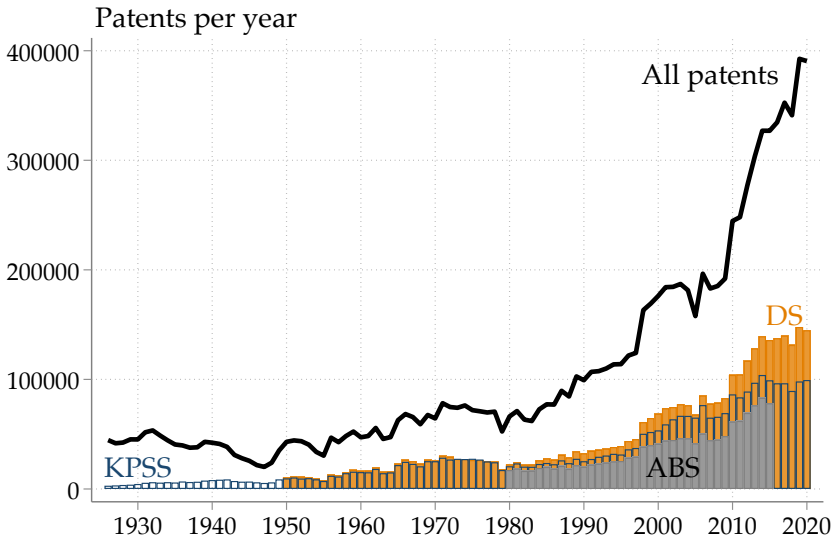


FIGURE 7. Comparison 2

Notes: This figure

- Number of patents
- Citation-weighted patents
- Probability of a blockbuster patent
- Scope of patent classes where
- Reliance on science

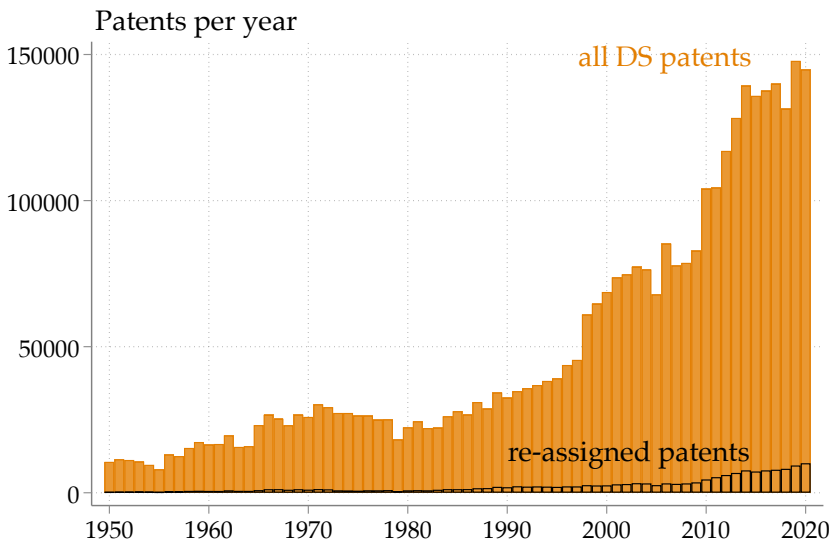


FIGURE 8. Comparison (re-assigned)

Notes: This figure

	Coverage	Dynamic	Firms	Patents	Disambiguated
DS 2023 Used in this paper	1950-2020	✓	**9,961** unique GVKEYs	3,115m	PatentsView + Harmonization with FGLMY + Extensive manual checks
ABS 2021	1980-2015	✓	4,985 unique PERMNOs	1,349m	Extensive manual checks
KPSS 2023	**1926-2023**	No	8,547 unique PERMNOs	**3,160m**	Some manual checks
KPSS 2023 Restricted to 1950-2020	1950-2020	No	8,448 unique PERMNOs	2.918m	Some manual checks
NBER 2001	1963-1999	No	2,487 unique CUSIPs	0.835m	Automatic

TABLE 9. Datasets of publicly-listed firms matched to patents

Notes: The numbers of patents and PERMNOs (unique firm identifier tied to a firm's stock) available in ABS 2021 are obtained from the `patent_1980_2015.dta` dataset from the authors (available here). The numbers for KPSS come from their `Match_patent_permco_permno_2022.csv` dataset (available here). The numbers for the NBER dataset come from the authors' `apat63_99.dta` dataset (available here).

The econometric specification takes the form:

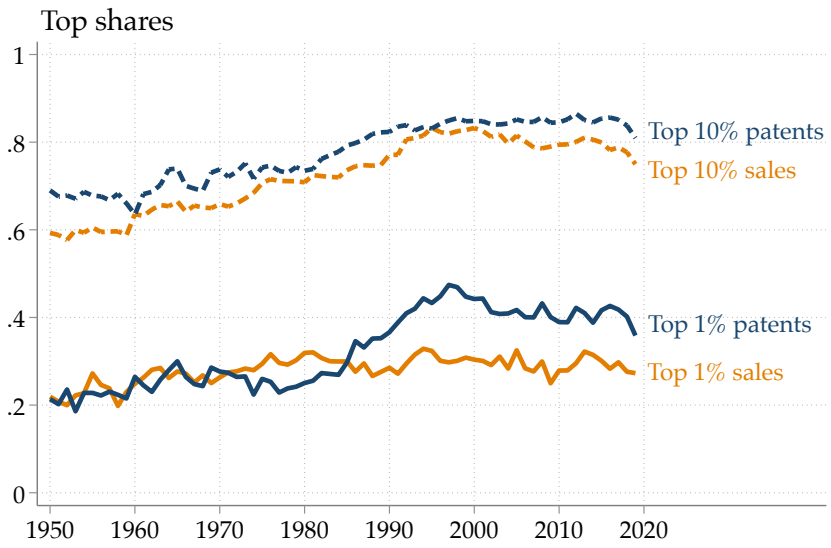


FIGURE 9. Top shares of sales and patents over time

Notes: This figure

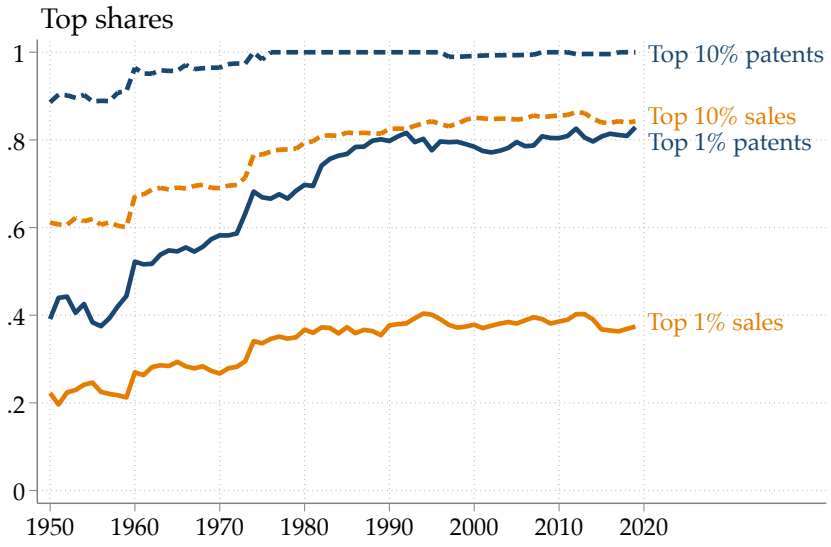


FIGURE 10. Top shares of sales and patents over time (all firms)

Notes: This figure

$$y_{it} = \alpha_i + \sum_t \beta_t \mathbb{1}[t]\mathbb{1}[\text{treat}] + \gamma[\text{treat}] + \text{year}_y + \mathbf{X}\beta + \varepsilon_{it}$$

5.4. **The relationship between firm scale and technology scope.**

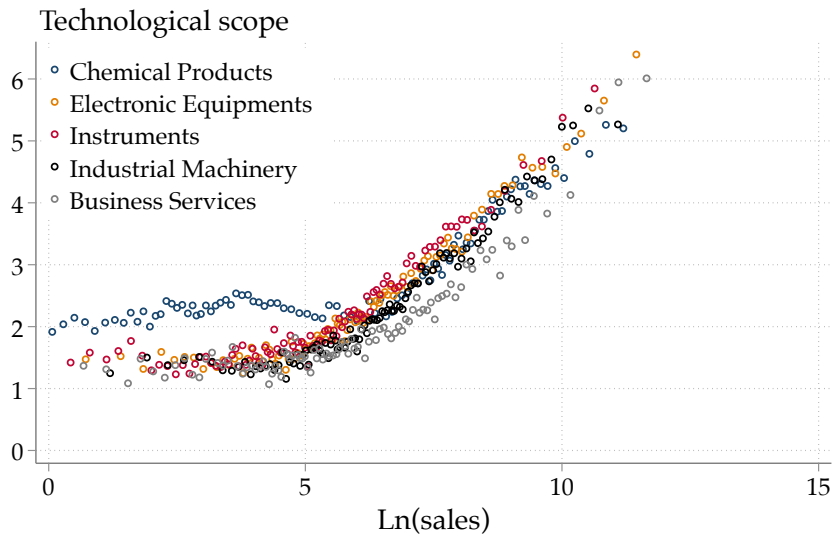


FIGURE 11. Scale and scope

Notes: This figure

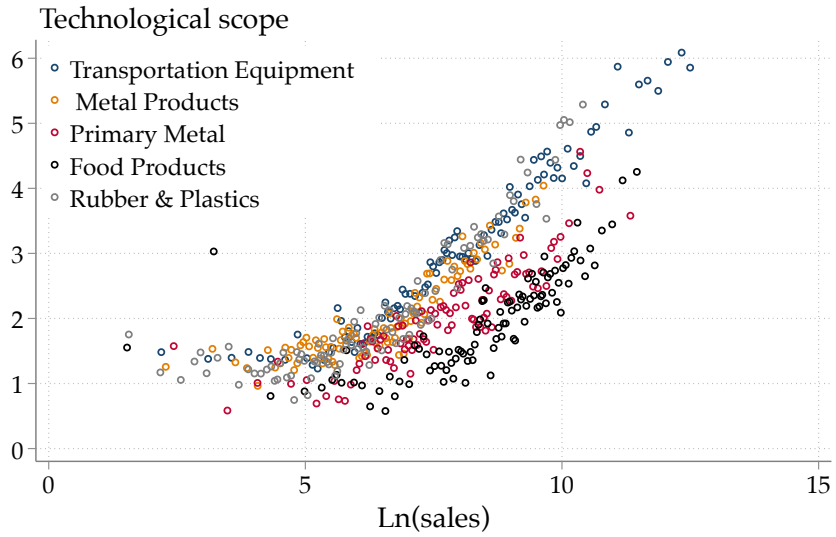


FIGURE 12. Scale and scope 2

Notes: This figure

6. CONCLUSION

This paper describes a newly assembled dataset of publicly listed firms from Compustat matched to patents over seven decades after World War II. We improve upon previous projects such as Arora et al. (2021b) and Hall et al. (2001) by doubling the period of the coverage; while previous datasets were only covering listed firms from 1980 to 2015, our datasets doubles the coverage and

	(1)	(2)	(3)	(4)
Dependent variable: ln(patent classes)				
ln(sale)	0.343***	0.197***	0.043***	0.035***
	(0.024)	(0.018)	(0.008)	(0.010)
ln(1 + patents)		0.510***	0.434***	0.597***
		(0.024)	(0.025)	(0.022)
ln(R&D)			0.264***	0.128***
			(0.018)	(0.015)
Year FE	✓	✓	✓	✓
Sector FE	✓	✓	✓	
Firm FE				✓
N	52,623	52,623	52,623	52,623
Adjusted R^2	0.480	0.663	0.695	0.846

TABLE 10. Firm scale and patent scope

Notes: The unit of analysis is a firm$t$×year. The table shows correlations between patent scope, defined as the log number of patent classes in which a firm files patents, and firm scale, defined here as log sales. Standard errors are two-way clustered at the year and SIC4 sector level.

extends it from 1950 to 2020. Our main contribution is to assemble a historical list of corporate events (such as M&As, relisting, name changes, etc.) to appropriately track how patents change hands over time. Overall, our dataset covers X% more firms than the second best, X% more patents and X% more corporate events. We hope this dataset can help researchers uncover new empirical regularities about corporate innovation in America over the long-run. Our dataset will be regularly updated, as new patent data and new firm data becomes available, as well as when we find inconsistencies within the existing data.

To demonstrate the potential of this dataset, we revisit a longstanding question of interest in the growth and innovation literature: how does innovativeness scale up with size, and did this relationship change over time. We find that X.

We believe this dataset can serve as an important step toward further developments of an even longer panel of public firms matched to patents. One natural extension of this dataset would be to use data in the annual reports of public firms before 1950[4], extract accounting data and merge it to pre-1950 patent data[5]

[4]For instance available through Mergent Archives

[5]Available from Fleming et al. (2019).

REFERENCES

AKCIGIT, U. AND W. R. KERR (2018): "Growth through heterogeneous innovations," *Journal of Political Economy*, 126, 1374–1443. [*Cited on pages 2 and 4.*]

ARORA, A., S. BELENZON, AND L. SHEER (2021a): "Knowledge spillovers and corporate investment in scientific research," *American Economic Review*, 111, 871–98. [*Cited on page 2.*]

——— (2021b): "Matching patents to compustat firms, 1980–2015: Dynamic reassignment, name changes, and ownership structures," *Research Policy*, 50, 104217. [*Cited on pages 3, 4, 9, 11, 12, and 21.*]

BERGEAUD, A., J. SCHMIDT, AND R. ZAGO (2022): "Patents that match your standards: firm-level evidence on competition and innovation," . [*Cited on page 2.*]

BESSEN, J. (2009): "NBER PDP project user documentation," *Data available at: https://sites. google. com/site/patentdataproject/Home/downloads*. [*Cited on pages 3 and 9.*]

BLOOM, N., M. SCHANKERMAN, AND J. VAN REENEN (2013): "Identifying technology spillovers and product market rivalry," *Econometrica*, 81, 1347–1393. [*Cited on page 2.*]

COHEN, W. M. (2010): "Fifty years of empirical studies of innovative activity and performance," *Handbook of the Economics of Innovation*, 1, 129–213. [*Cited on pages 3 and 4.*]

FLEMING, L., H. GREENE, G. LI, M. MARX, AND D. YAO (2019): "Government-funded research increasingly fuels innovation," *Science*, 364, 1139–1141. [*Cited on pages 2, 4, 8, and 22.*]

HALL, B. H., A. B. JAFFE, AND M. TRAJTENBERG (2001): "The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools," Working Paper 8498, National Bureau of Economic Research. [*Cited on pages 3 and 21.*]

APPENDIX A. APPENDIX SECTION