Knowledge Diffusion and R&D Cost Spillovers in U.S. Medical Innovation: A Network Approach

Michele Liberatore* Massimo Riccaboni†

This version: November 2026

Abstract

We study whether R&D expenditures generate cost spillovers across technologically related pharmaceutical projects. We link the costs of clinical trials registered on ClinicalTrials.gov to a citation network of the U.S. patents associated with them, and estimate spatial econometric models in which the cost of an invention depends on the R&D embodied in the knowledge it builds on. To address endogenous network formation, we measure the technological similarity between every pair of patents from the full text of their titles, abstracts and claims, and decompose the citation network into components that separate co-movement due to similarity from co-movement along citations. Although technological similarity accounts for most of the co-movement of costs among related inventions, a citation-specific channel survives this control, consistent with a genuine cost spillover between patents. As for its direction, exposure to costlier prior R&D is associated with higher development costs: a 10% increase in the R&D embodied in the knowledge an invention builds on goes along with a 1.2% higher cost. The relationship is twice as strong for knowledge produced within the owning organization, but extends across organizational boundaries. Moreover, it originates from upstream successes, whereas the costly failures of other organizations are associated with lower costs of later development. Our results point to a crowding effect of prior R&D investments on the cost of new development, and suggest that the measurement of drug development costs should extend beyond the boundaries of the developing firm.

Keywords: *Knowledge diffusion; R&D costs; Pharmaceutical innovation; Clinical trial outcomes; Patent citation networks; Endogenous network formation*
JEL codes: O31; O33; L65; I11; C21

1 Introduction

Bringing a new drug to market requires investments that now amount to several billion dollars per approved compound (DiMasi et al., 2016; Wouters et al., 2020), and the length and cost of clinical trials shape which projects firms choose to pursue (Budish et al., 2015). A long tradition in health economics has shown that a large share of this figure reflects the cost of the projects that fail along the way, and that the knowledge generated by successes and failures can affect the performance of later projects developed by the same firm (Magazzini et al., 2012; Chiou et al., 2016). In this accounting, each developer is treated as a self-contained unit: the cost of a new drug includes the direct and indirect costs borne by the firm that develops it, but not the costs of the related projects carried out by others in the same technological field. Yet pharmaceutical research is highly cumulative. New compounds build on the mechanisms, molecules and clinical

*IMT School for Advanced Studies Lucca, Italy and University of Florence, Italy. Email: michele.liberatore@imtlucca.it

†IMT School for Advanced Studies Lucca, Italy and IUSS Pavia, Italy. Email: massimo.riccaboni@imtlucca.it

evidence produced by earlier projects, many of which are developed elsewhere. If the knowledge produced by others affects the resources that a firm must invest in its own development, the private accounting of R&D costs misses an externality that is central to both the economics of innovation and the design of pharmaceutical policy.

An extensive literature has studied knowledge spillovers through a variety of outcomes: patenting and productivity (Jaffe, 1986; Bloom et al., 2013; Lucking et al., 2019), the research productivity of drug discovery (Henderson and Cockburn, 1996), the private patenting induced by public biomedical research (Azoulay et al., 2019), the output of scientists and publicly funded research (Zacchia, 2020; Myers and Lanahan, 2022), and the continuation and success of clinical projects (Krieger, 2021; Chiou et al., 2016). Almost all of this evidence concerns the output side of innovation. Much less is known about the input side, that is, whether the knowledge produced by others makes the development of a new project cheaper or more expensive. This paper addresses this question by asking whether R&D expenditures generate cost spillovers across technologically related pharmaceutical projects, and whether these cost spillovers depend on how the upstream projects turned out.

Theory offers two opposing predictions. Under a *standing-on-shoulders* view, drawing on a larger and costlier body of prior research lowers the cost of new development: earlier projects resolve scientific uncertainty, inform the design of later trials and allow developers to avoid unpromising paths (Scotchmer, 1991). Under a *crowding* view, heavy prior investment marks a dense and well-explored segment of the technology space, in which a marginal contribution requires larger trials, must outperform a growing set of existing therapies and competes with parallel efforts for the same resources (Jones, 2009; Bloom et al., 2020; Scannell et al., 2012). The two views also differ in what they imply for the outcome of upstream projects. A successful programme provides validated targets and designs that later projects can reuse, but also an effective therapy that they must beat; a failed programme reveals which paths to avoid, but may also be followed by projects that repeat its design. The net effect of prior knowledge on development costs is therefore ambiguous, and its sign, for successful and failed upstream projects alike, remains an open empirical question.

To answer it, we combine trial-level cost estimates for drug and biologic trials registered on ClinicalTrials.gov with a network of citations among the U.S. patents associated with those trials, and estimate spatial econometric models in which the cost of an invention depends on the R&D embodied in the knowledge it builds on and co-moves with the costs of related inventions. A central difficulty is that citations are not formed at random: inventions cite technologically similar prior art, and similar inventions have similar costs for reasons unrelated to knowledge flows. We address this endogenous network formation by measuring the technological similarity between every pair of patents from the full text of their titles, abstracts and claims, using both transformer-based embeddings and keyword vectors, and by decomposing the citation network into disjoint components that separate co-movement due to similarity from co-movement that travels along citations. We then use the development histories of the programmes behind each patent to sort the knowledge an invention builds on by whether, at the time, it was known to come from a successful or a failed project, and by whether it was produced by the same owner or by others.

Our results support the crowding view. Although technological similarity accounts for most of the co-movement of costs among related inventions, a citation-specific channel survives the control for similarity, and exposure to costlier prior R&D is associated with higher development costs: a 10% increase in the R&D embodied in the knowledge an invention builds on goes along with a 1.2% higher cost, an association driven by the knowledge on which the cited inventions were themselves built. The association is about twice as strong for knowledge produced within the owning organization, but remains positive and significant across organizational boundaries, and it is robust to alternative definitions and measures of similarity. More specifically, it originates from upstream successes, as the R&D embodied in cited successes is associated with

1.3–1.5% higher costs per 10% increase, while cited failures show no association on average. This average, however, hides two opposite patterns: costly failures of the same owner are associated with higher later costs, whereas the costly failures of other organizations are associated with lower ones, consistent with developers learning from the failures of others.

The paper contributes to four strands of research. First, it shifts the study of knowledge spillovers from the output to the input side of innovation, providing evidence on cost spillovers across pharmaceutical projects. Second, it extends the accounting of drug development costs beyond the boundaries of the developing firm, showing that the costs of related projects carried out by others feed into the cost of own development. Third, it relates the value of upstream knowledge to the outcome of the projects that produced it, complementing the evidence on learning from failures within firms (Magazzini et al., 2012; Chiou et al., 2016) and across firms (Krieger, 2021) with evidence on development costs. Fourth, it proposes a strategy to address endogenous network formation in citation networks, based on text-based similarity and on the decomposition of the network into disjoint components (Castellacci et al., 2026; Arts et al., 2023; Ghosh et al., 2024).

The remainder of the paper is organized as follows. Section 2 reviews the related literature, and Section 3 sets out the conceptual framework. Section 4 describes the data, and Section 5 the construction of the knowledge networks. Section 6 presents the empirical model, Section 7 the results, and Section 8 discusses them. Section 9 concludes.

2 Related literature

2.1 The cost of drug development and the value of upstream outcomes

A long tradition in health and industrial economics measures the resources required to bring a new drug to the market. The canonical approach, developed by DiMasi et al. (1991, 2003, 2016), aggregates the out-of-pocket cost of clinical trials across phases, accounts for the probability of attrition and capitalizes expenditures over the development horizon. Overall, its central insight is that the cost of an approved drug must also include the cost of the projects that failed along the way. Under this accounting scheme, the capitalized cost per approved compound has risen to several billion dollars. Estimates based on public filings (Wouters et al., 2020) and on trial-level cost models (Sertkaya et al., 2024) arrive at lower but still substantial figures, and confirm that the costs of failure and of capital account for the largest share of the total. These costs are not neutral for the direction of research: the length and expense of clinical trials shape which projects firms choose to undertake (Budish et al., 2015).

However, failed projects are not merely sunk costs but also inputs into later research. Patents associated with failed pharmaceutical projects continue to be cited by later inventions, and firms can learn from both their own successes and their own failures, although the two types of experience affect subsequent performance differently (Magazzini et al., 2012; Chiou et al., 2016). The information produced by a clinical outcome also travels across firm boundaries. Firms respond to the failures of competitors that share their market and technology by abandoning their own projects, suggesting that technological learning outweighs the gains from reduced competition (Krieger, 2021). Since novel candidates face higher risk but promise greater value (Krieger et al., 2022), the knowledge produced by upstream outcomes is likely to be most informative precisely where development is most uncertain. This information has also become increasingly public since the FDA Amendments Act of 2007 made trial registration and the reporting of summary results on ClinicalTrials.gov mandatory for most trials of regulated products (Zarin et al., 2011). Whether an upstream programme advanced to the next phase, reached approval or was discontinued is therefore an observable signal that later developers can act upon.

What this literature has not established is whether the R&D expenditure of one project affects the *cost* of later, technologically related projects undertaken elsewhere. Within-firm learning implies that a firm's past outlays can reduce its future costs. Whether exposure to the

costly knowledge produced by others reduces or raises the cost of own development, and whether this depends on how the upstream projects turned out, remains an open empirical question. This is the gap we address.

2.2 Knowledge spillovers, patent citations and the burden of knowledge

Since Griliches (1979) and Jaffe (1986), the economics of R&D spillovers has measured the external knowledge available to a firm by weighting the R&D of others with a measure of technological proximity. Bloom et al. (2013) separate technology spillovers from product-market rivalry and find that the social returns to R&D substantially exceed the private returns. Later work studies specific channels of diffusion: networks of scientists (Zacchia, 2020), publicly funded research (Azoulay et al., 2019; Myers and Lanahan, 2022) and markets for technology, which internalize part of the flows conventionally treated as externalities (Arqué-Castells and Spulber, 2022). In pharmaceuticals, Henderson and Cockburn (1996) show that spillovers from rivals' research raise the productivity of drug discovery, and network analyses of the industry document how the division of innovative labour among firms shapes the diffusion of biomedical knowledge (Orsenigo et al., 2001).

Patent citations have been the standard empirical trace of knowledge flows since Jaffe et al. (1993), and citation networks are increasingly used to map how innovation in one area builds on innovation in another. Acemoglu et al. (2016) show that past innovation in upstream technology classes predicts future innovation downstream, and Liu and Ma (2021) use a global citation network to study the allocation of R&D across sectors. Citations are nonetheless a noisy measure of knowledge use. Because applicants are subject to a duty to disclose known material prior art, citations are numerous and partly defensive; a share is added by examiners rather than inventors (Alcácer and Gittelman, 2006); the number of citations per patent has grown sharply over time (Kuhn et al., 2020); and citations are more likely between technologically close inventions irrespective of any knowledge flow. A recent literature therefore measures proximity directly from patent text, using keyword overlap (Arts et al., 2018, 2021, 2023), text-based measures of novelty and impact (Kelly et al., 2021) or transformer embeddings trained on patent documents (Ghosh et al., 2024). We use these measures not as a substitute for citations but as a control for the part of the citation network that reflects technological proximity.

Almost all of this literature studies spillovers on the output side of innovation, through productivity, patenting or market value. Evidence on the input side, that is, on R&D costs, is scarce, and theory offers two opposing predictions. The cumulative-innovation view (Scotchmer, 1991) implies that building on a larger body of prior work lowers the cost of new development, a *standing-on-shoulders* effect. A second view stresses that knowledge accumulation raises the burden faced by later innovators (Jones, 2009), that new ideas become progressively harder to find (Bloom et al., 2020), and that in pharmaceuticals each new drug must outperform an expanding set of existing therapies (Scannell et al., 2012; Pammolli et al., 2011). Under this *crowding* view, a marginal contribution to a dense and well-explored segment requires larger investments. Prior intellectual property may reinforce this channel, since follow-on projects in densely patented areas must license, invent around or litigate upstream claims (Heller and Eisenberg, 1998; Williams, 2013; Sampat and Williams, 2019). We bring these predictions to the data by asking whether exposure to costlier prior knowledge raises or lowers the cost of current R&D, and whether the answer differs for knowledge that comes from successful and from failed upstream programmes.

2.3 Peer effects and endogenous network formation

Estimating spillovers on a network raises two distinct identification problems. The first is simultaneity. When outcomes are jointly determined across connected units, the spillover term is endogenous (Manski, 1993), although peer effects can be identified from the structure of the

network itself under suitable conditions (Bramoullé et al., 2009). Spatial autoregressive models address this problem by modelling the full feedback structure and decomposing effects into direct and indirect components (LeSage and Pace, 2009). König et al. (2019) and Hsieh et al. (2025) derive such specifications as the reduced form of equilibrium models of R&D collaboration, and Bayesian estimators extend them to large, unbalanced panels with time-varying networks (Grieser et al., 2022). Castellacci et al. (2026) apply this approach to publicly funded R&D consortia and show that spillover pools built on the network of collaborators raise innovation performance.

The second problem is that links are not formed at random. If similar units are more likely to connect and also share unobserved determinants of the outcome, homophily in link formation is confounded with peer effects (Goldsmith-Pinkham and Imbens, 2013; Johnsson and Moon, 2021). The problem is particularly acute in citation networks: an invention cites prior art because the two are technologically similar, and similar projects are likely to have similar development costs for reasons unrelated to any knowledge flow. Our approach combines the two strands. We estimate a spatial Durbin model on a directed, time-varying network of patent citations, and use text-based similarity, measured for every pair of patents, to separate the links that reflect technological proximity from those that do not.

3 Conceptual framework

3.1 A cost equation with knowledge exposure

We model the cost of a pharmaceutical project as a function of the R&D embodied in the prior knowledge it builds on, and allow it to co-move with the costs of the projects it is connected to. Consider a set of inventions, indexed by i, each embodied in a patent and developed through one or more clinical trials. Let c_{it} denote the (log) R&D cost accumulated by invention i up to period t. Inventions are connected by directed knowledge links. Let $W_t = [w_{ijt}]$ be the corresponding weight matrix, with $w_{ijt} > 0$ if invention i builds on invention j by period t, $w_{iit} = 0$, and rows normalized to sum to one for every invention with at least one link. We define the knowledge exposure of invention i, x_{it}, as the (log) R&D cost embodied in the inventions it builds on. Exposure therefore measures the amount of costly prior knowledge available to the project and, by construction, is determined by investments that precede the project itself.

We posit the following reduced-form cost equation:

$$c_{it} = \rho \sum_j w_{ijt}\, c_{jt} + \beta\, x_{it} + \theta \sum_j w_{ijt}\, x_{jt} + Z_{it}\gamma + \alpha_{g(i),t} + \varepsilon_{it}, \tag{1}$$

where Z_{it} collects invention characteristics and $\alpha_{g(i),t}$ absorbs shocks common to the technological field $g(i)$ in period t. The equation contains two distinct network terms. The first, governed by ρ, relates the cost of an invention to the contemporaneous cost of the inventions it is linked to. It captures the co-movement of development costs along knowledge links and is endogenous, since costs are jointly determined across connected projects. The second, governed by θ, relates the cost of an invention to the exposure of its neighbours, that is, to the R&D embodied in the knowledge on which its own sources were built. Because exposure is predetermined, θ captures a flow from upstream R&D to the cost of the current project rather than the joint determination of two outcomes. Together with the own-exposure coefficient β, it summarizes how costly prior knowledge feeds into current R&D expenditure. In other words, the two exposure terms look one and two steps upstream along the citation network, respectively: x_{it} measures the cost of the inventions that i cites, and $\sum_j w_{ijt}x_{jt}$ the cost of the knowledge on which those inventions were built. Equation (1) is a spatial Durbin specification. It nests the spatial autoregressive model ($\theta = 0$) and the spatial lag of X model ($\rho = 0$).

The two hypotheses of Section 2 map into the sign of the effect of exposure on cost. Under the *standing-on-shoulders* hypothesis, drawing on a larger body of costly prior research reduces

the cost of new development. Prior trials resolve scientific uncertainty, inform the design of subsequent studies and allow later projects to avoid unpromising paths, so exposure lowers cost. Under the *crowding* hypothesis, heavy prior investment marks a dense and well-explored segment of the technology space. Adding a marginal contribution there requires clearing a higher evidentiary bar, demonstrating superiority over a larger set of existing alternatives and competing for the same patients and trial sites, so exposure raises cost. Equation (1) does not impose either mechanism. The data determine which dominates on average.

3.2 Direct, indirect and total effects

Because the outcome appears on both sides of equation (1), the effect of exposure is not given by a single coefficient. Stacking observations for period t and solving for the reduced form yields

$$c_t = (I - \rho W_t)^{-1} (\beta I + \theta W_t) x_t + (I - \rho W_t)^{-1} (Z_t \gamma + \alpha_t + \varepsilon_t). \tag{2}$$

The matrix $S_t = (I-\rho W_t)^{-1}(\beta I+\theta W_t)$ collects the partial effects of exposure on cost: its element (i, j) measures the effect of a change in the exposure of invention j on the cost of invention i, once all feedback through the network is accounted for. The term $(I-\rho W_t)^{-1} = I+\rho W_t+\rho^2 W_t^2+\ldots$ is the network's Leontief multiplier. A change in the cost of one project shifts the costs of the projects linked to it, which in turn shift the costs of their own neighbours, and so on along every path of the network. Following LeSage and Pace (2009), we summarize S_t by three scalar measures. The *direct* effect is the average diagonal element of S_t: the effect of an invention's own exposure on its own cost, including the feedback that returns to it through the network. The *total* effect is the average row sum of S_t: the effect on an invention's cost of a uniform increase in the exposure of all inventions. The *indirect* effect is the difference between the two, and measures the part of the effect that operates through the exposure of other inventions. It is the network spillover of upstream R&D.

With a row-normalized W_t, these measures take a simple form. Since every non-zero row of W_t sums to one, the total effect is approximately

$$\left.\frac{\partial c}{\partial x}\right|_{\text{total}} \approx \frac{\beta + \theta}{1 - \rho}, \tag{3}$$

that is, the immediate effect of exposure, $\beta+\theta$, amplified by the network multiplier $1/(1-\rho)$. The direct effect is approximately β, and the indirect effect approximately $\theta/(1-\rho)$. Equation (3) is the quantity that discriminates between the two hypotheses: a negative total effect supports the standing-on-shoulders view, whereas a positive total effect supports the crowding view. The coefficient ρ does not create a spillover of exposure by itself; it determines how strongly the spillovers generated by β and θ are amplified as they travel through the network.

3.3 Endogenous network formation

The interpretation above presumes that the links in W_t are channels of knowledge diffusion. In practice, an invention builds on prior work that is technologically close to it, and technologically close projects are likely to have similar costs for reasons that have nothing to do with knowledge flows: they target the same therapeutic areas, involve similar molecules and face similar regulatory requirements. If links form between similar inventions, the co-movement of costs captured by ρ may reflect this assortativity rather than diffusion, and θ may partly absorb it.

To separate the two, suppose that pairwise technological similarity between inventions can be measured independently of the links. Every pair can then be assigned to one of three disjoint sets: pairs that are linked but not similar (W_t^c), pairs that are linked and similar (W_t^b), and pairs that are similar but not linked (W_t^s). Replacing the single peer term in equation (1) with

$$\rho_1 \sum_j w_{ijt}^b \, c_{jt} + \rho_2 \sum_j w_{ijt}^s \, c_{jt} + \rho_3 \sum_j w_{ijt}^c \, c_{jt} \tag{4}$$

yields testable predictions. If cost co-movement were driven entirely by assortativity, it would arise among similar pairs regardless of whether they are linked, so that $\rho_1 \approx \rho_2$ and $\rho_3 = 0$. If, instead, knowledge links carry information on costs beyond technological proximity, co-movement should also be present among linked pairs that are not similar, so that $\rho_3 > 0$. The coefficient ρ_3 is therefore the component of network co-movement that assortativity cannot explain, while ρ_2 measures assortativity in its pure form. The exposure terms β and θ, which are built from predetermined upstream R&D, remain the measures of the spillover of interest; the decomposition protects them from absorbing contemporaneous co-movement among similar projects.

3.4 Upstream outcomes: successes and failures

The framework above treats all upstream R&D alike. Yet the knowledge produced by a costly programme that advanced through development differs from the knowledge produced by a costly programme that was discontinued, and the two hypotheses of Section 3.1 make different predictions for each. Because clinical outcomes become public once trials are completed, a later developer can condition on whether the projects it builds on succeeded or failed, and the exposure terms of equation (1) can be sorted accordingly, with the cost of each cited invention attributed to the outcomes of its own projects as known at the time.

Successful upstream programmes produce validated targets, doses, endpoints and trial designs that later projects can reuse, which lowers their cost under the standing-on-shoulders view. At the same time, a success marks a field in which an effective therapy already exists, so later entrants must demonstrate an advantage over it and recruit from a patient population that is already served, which raises their cost under the crowding view. Failed programmes, in turn, reveal which paths are unpromising and allow later projects to avoid them, a channel through which costly failures should lower subsequent costs (Krieger, 2021). They may instead raise costs when later projects repeat the designs, comparators and indications of programmes that did not work. The prevalence of each channel is likely to depend on whether the upstream knowledge was produced by the same organization. An owner has privileged access to the information generated by its own programmes, but may also carry forward their cost structure, while it observes the outcomes of others' programmes through public disclosure and builds on them selectively. For each outcome, the quantity that discriminates between the hypotheses is the analogue of equation (3): the combined total effect of exposure to cited inventions of that outcome, distinguishing internal from external knowledge.

4 Data

4.1 Data sources

Our empirical analysis combines five data sources, which link U.S. pharmaceutical patents to the clinical trials in which the underlying inventions are developed, to the characteristics, costs and outcomes of those trials, and to the bibliographic and textual characteristics of the patents themselves. Table 1 summarizes the information drawn from each source.

Three features of the data deserve mention. First, the trial–patent links in PKG 2.0 are citation-based: a trial registered on ClinicalTrials.gov is connected to the patents that are scientifically and technologically related to it, not only to the patent protecting the commercialized product. This makes PKG 2.0 the natural basis for a network of knowledge flows. The relationship is many-to-many, as a trial typically involves several patents and a patent may be associated with several trials. All patents in the sample are granted by the United States Patent and Trademark Office (USPTO). Second, CTO assigns each drug and biologic trial in Phases 1–3 an outcome label aggregated from several weakly supervised signals; where a trial is registered under a combined phase, we retain the highest phase reached. We use these labels to

Table 1: Data sources

Source	Information used	Coverage
PKG 2.0 (Xu et al., 2025)	Trial–patent links	27,062 pairs; 4,747 trials; 6,790 patents
ClinicalTrials.gov	Primary completion date	4,747 trials
CTO (Gao et al., 2024)	Trial phase and outcome	3,542 labelled trials (Phases 1–3)
Evaluate Pharma	Trial cost ($ million), start date, programme phase history and status	2,747 trials with cost
Orbis IP	Citations and citation type, priority and filing dates, CPC codes, family size, applicants and owners, title, abstract and claims	4,635 USPTO patents

Notes: Coverage refers to the trials and patents relevant to our sample. PKG 2.0: PubMed Knowledge Graph 2.0. CTO: Clinical Trial Outcome benchmark. Trials are identified across sources by their ClinicalTrials.gov registration number.

validate the outcome measure built from Evaluate (Section 4.4). Third, Evaluate's trial costs are model-based estimates: each trial's cost is obtained by applying to its enrollment and duration a cost-per-patient benchmark, computed from the R&D spend reported in 10-K filings over 2007–2025 and pooled within technology and therapeutic class, with fixed adjustments for trial location. Evaluate also records the development history of each programme, that is, of each product developed for a given indication, from which we build the outcome measure that CTO labels validate.

4.2 Sample construction

Figure 1 summarizes the construction of the sample. On the trial side, the estimation sample consists of trials that appear in all three trial-level sources: linked to at least one patent in PKG 2.0, labelled by CTO, and with a cost estimate in Evaluate. Of the 4,747 patent-linked trials, 3,542 (74.6%) are covered by CTO; the remainder fall outside its coverage of drug and biologic trials in Phases 1–3 (e.g. device or Phase 4 studies). All covered trials are found in Evaluate, and 2,747 of them have a cost figure attached. Excluding 174 trials without a start date in Evaluate, which is required to place each trial in time, leaves 2,573 trials. On the patent side, the sample follows the selection of trials: of the 6,790 patents in the PKG 2.0 links, 4,635 are linked to the retained trials, and 4,574 of them have a priority date, which places the patent in time. Dropping the 15 trials whose linked patents all lack a priority date, the final sample contains 2,558 trials, 4,574 patents and 13,605 citation links among them. Almost all sample trials (99.6%) have concluded, so that their costs are not truncated by ongoing development.

4.3 Descriptive statistics

Table 2 describes the distribution of trial costs in the estimation sample. Costs are strongly right-skewed: the median trial costs $14.6 million while the mean exceeds $50 million. Cost rises steeply with phase, from a median of $6.1 million in Phase 1 to $12.9 million in Phase 2 and $49.5 million in Phase 3, and Phase 3 trials account for about two thirds of total expenditure in the sample.

Figure 2 shows how the sample is distributed over time. Trial starts rise steadily from the early 2000s, peak in the mid-2010s and decline thereafter, as more recent trials are less likely to have been costed and linked to patents (panel a). Patent priority years follow the same

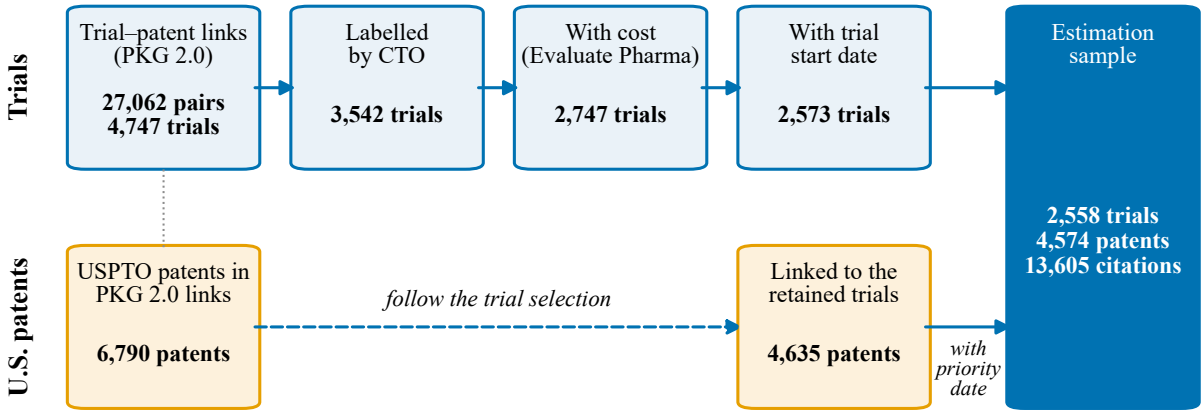


Figure 1: Construction of the sample

Notes: Top row: selection of clinical trials registered on ClinicalTrials.gov. Bottom row: U.S. patents linked to those trials in PKG 2.0. The estimation sample retains the trials and patents that satisfy all conditions and the citation links among the retained patents.

Table 2: Trial costs by phase ($ million)

Phase	Trials	Median	Mean	Total
Phase 1	733	6.1	17.3	12,659
Phase 2	1,169	12.9	27.5	32,189
Phase 3	654	49.5	128.1	83,758
All	2,558	14.6	50.3	128,709

Notes: Trials in the estimation sample. Trials registered under a combined phase (Phase 1/2 or Phase 2/3) are assigned to the higher of the two phases.

pattern with a lead of a few years, consistent with inventions being patented before their clinical development begins, and the cumulative number of citation links among sample patents grows until the late 2010s, when the entry of new citing patents slows down (panel b). On the patent side, the many-to-many structure of the data is substantial: the average trial is linked to about six patents, and the average patent is associated with about 3.4 trials. A large share of patents is therefore shared across several development projects, which motivates the allocation of trial costs to patents described in Section 5.1.

4.4 Outcomes of development programmes

To measure whether upstream projects succeeded or failed, we exploit the development histories recorded by Evaluate. Evaluate tracks each programme, that is, each product developed for a given indication, through the phases of development, and records its current status (phase reached, approved, or discontinued) together with the programme, phase and start date of each of its trials. We match 98% of sample trials to their programme. Each trial is then labelled by what happened to its programme after the trial started, with respect to the next phase of development (Phase 1 to Phase 2, Phase 2 to Phase 3, Phase 3 to regulatory filing or approval). A trial is *advanced* if its programme later reached the next phase, and *failed* if the programme was discontinued without reaching it or remained at that phase for more than five years. Trials started after the programme had already reached a later phase, such as post-approval studies, and trials too recent for their outcome to be determined are left undecided.

This measure has three properties that matter for what follows. First, the label refers to

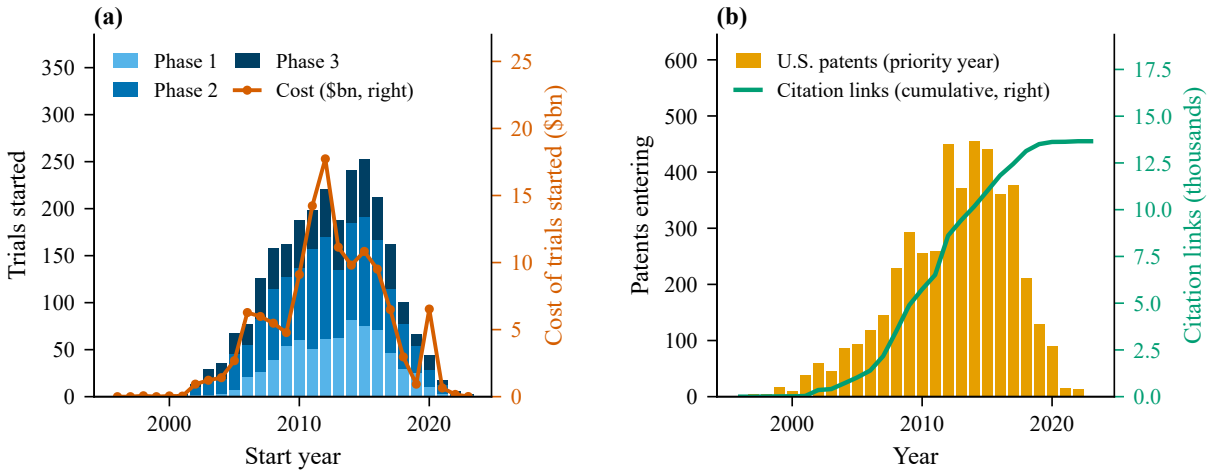


Figure 2: Trials, costs, patents and citation links over time

Notes: Panel (a): number of sample trials by start year and phase (left axis) and total cost of the trials started in each year, in $ billion (right axis). Panel (b): number of sample U.S. patents by priority year (left axis) and cumulative number of citation links among sample patents, dated at the priority year of the citing patent (right axis).

the programme rather than to the individual trial: a trial may meet its primary endpoint and yet belong to a programme that is later discontinued for commercial or strategic reasons. What matters for later developers, however, is precisely whether the line of development continued, and this is the information conveyed by the programme history. Second, outcomes are dated. An outcome is considered known from the year after the trial's primary completion date as registered on ClinicalTrials.gov, and a programme that stalls counts as a known failure only five years after the trial's start. A later invention can therefore respond only to the outcomes revealed before it was developed (Section 3.4). Third, the programme-based labels are consistent with an independent, trial-level source where the two definitions coincide. CTO also defines success as progression to the next phase, but infers it from trial-level signals rather than from programme histories. Among the 2,117 sample trials that carry both labels, 93% of the trials whose programme advanced are also classified as successes by CTO, and 72% of the trials classified as failures by CTO belong to programmes that failed. The two sources diverge mainly on trials that CTO records as successful but whose programme was later discontinued or stalled. These are the cases that a trial-level label cannot capture and a programme history can, which is why we rely on the latter.

5 Construction of the knowledge networks

Using the data described above, we build two distinct objects on the set of U.S. patents in our sample. The first is a directed, time-varying citation matrix, which records the knowledge flows between inventions and on which we define the cost and exposure variables used in the estimation. The second is a similarity matrix, which records how technologically close every pair of inventions is, irrespective of whether one builds on the other. We then combine the two to partition the citation network into disjoint components.

5.1 The citation matrix

Each node of the network is a distinct patent associated with at least one trial in the sample. A directed link from patent j to patent i is established if i cites j and both patents belong to the sample, while citations to patents outside the sample are discarded. Links are oriented in the direction of knowledge flow, from the cited (earlier) to the citing (later) invention. The resulting

network contains 13,605 citation links among the 4,574 patents. Applicant-added citations account for 93% of all links, the remainder being added by examiners or search authorities, which mitigates the concern that citations reflect bureaucratic prior-art insertion rather than knowledge actually used by inventors (Alcácer and Gittelman, 2006).

A citation is a dated event, since it can only be made by an existing invention to an invention that precedes it. We therefore place each link in time at the priority date of the citing patent, so that the link enters the network in the period in which the citing invention is conceived and remains active thereafter, while each patent enters the panel in the period of its own priority date. Periods are years, from 1996 to 2023, which yields a sequence of 28 citation matrices W_t that grows monotonically as new inventions enter and build on earlier ones. The orientation of links is consistent with calendar time, as the citing patent is filed on average about two years after the cited one (median gap of 1.2 years). Because a later invention can cite an earlier one but not vice versa, the citation network is a directed acyclic graph: when patents are ordered by entry, W_t is triangular and there is no mutual feedback between a citing and a cited invention. This ordering ensures that knowledge exposure, which is built from the cited inventions only, is predetermined with respect to the cost of the citing one. In the baseline specification, links are binary and W_t is row-normalized, so that each active link of patent i receives weight $w_{ijt} = 1/n_{it}$, where n_{it} is the number of patents i cites that are active in period t. Network terms are therefore averages over a patent's neighbours, which gives every neighbourhood the same total weight, prevents the estimates from being driven by a small number of heavily connected patents, and makes coefficients comparable across networks of different density.

The cost and exposure variables are defined on this network. Since R&D cost is observed at the trial level while citations live at the patent level, we allocate the cost of each trial equally across the patents it is linked to, and define the cost of a patent as the sum of its shares:

$$C_i = \sum_{k \in T(i)} \frac{C_k}{|P(k)|}, \tag{5}$$

where $T(i)$ is the set of trials associated with patent i, C_k is the cost of trial k, and $|P(k)|$ is the number of patents linked to trial k. This fractional allocation ensures that no monetary unit is counted more than once, as allocated costs summed over all patents exactly reproduce total trial cost. Each trial's share is booked in the period in which the trial starts, and the dependent variable c_{it} is the logarithm of one plus the cumulative cost of patent i up to period t. The knowledge exposure of patent i in period t is the cumulative cost of the patents it cites through links active by t:

$$x_{it} = \ln\Big(1 + \sum_{j:\, w_{ijt} > 0} \tilde{C}_{j,-i}\Big), \tag{6}$$

where $\tilde{C}_{j,-i}$ is the cost of the cited patent j net of any share deriving from trials in which patent i itself is involved. This leave-one-out correction is necessary because a trial shared by the citing and the cited patent would otherwise enter both the dependent variable and the regressor, so that exposure is built exclusively from R&D external to the focal patent.

Figure 3 describes the resulting network and the distribution of patent costs. Citation links are unevenly distributed (panel a). The median patent receives four citations from other sample patents and makes three, but a small group of patents is cited far more often, up to 89 times, so that the top 1% of cited patents receive about 10% of all citations. On one hand, citations made are more bounded, as no patent cites more than 62 others in the sample. On the other hand, patent costs are approximately log-normal, with a median of \$4.4 million and a mean of \$27.9 million, and span five orders of magnitude (panel b).

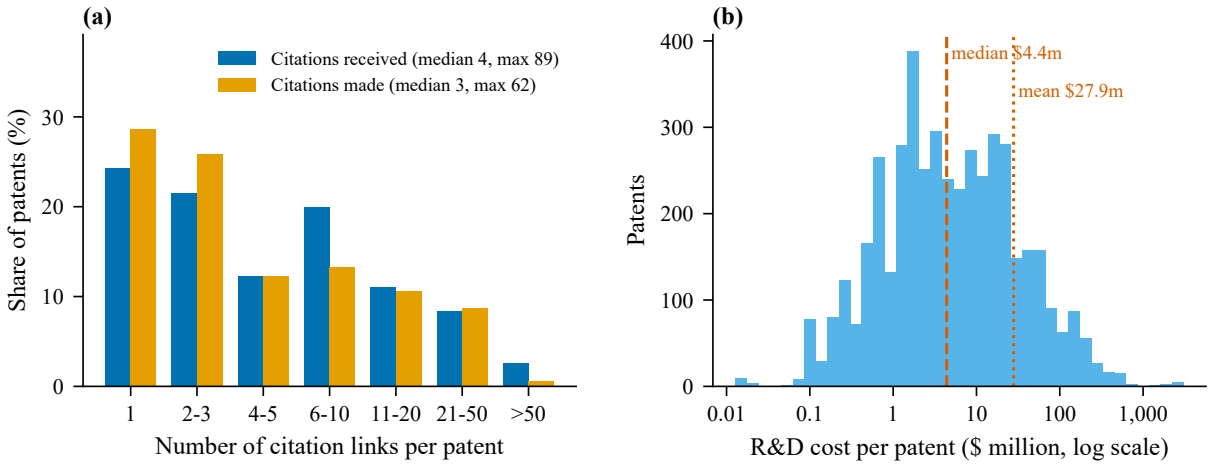


Figure 3: Citation links and R&D cost per patent

Notes: Panel (a): share of sample patents by number of citations received from, and made to, other sample patents; patents with no link of the given type are excluded. Panel (b): distribution of the cumulative R&D cost allocated to each patent according to equation (5), on a logarithmic scale; dashed line: median; dotted line: mean.

5.2 The similarity matrix

To address endogenous network formation (Section 3.3), we measure the technological similarity between every pair of patents from the full text of their titles, abstracts and claims, and collect it in a similarity matrix $S = [s_{ij}]$. Claims are central to this choice: unlike abstracts, which are short and often generic, claims define precisely what is legally protected, namely the class of compounds, their substituents, the formulation and the therapeutic use, and are therefore the most precise textual description of an invention's content (Arts et al., 2023). Text is available for all sample patents, the median patent has 17 claims corresponding to about 600 words, and all markup in the source text is removed before processing. The resulting matrix differs from the citation matrix in three aspects. First, it is defined for every pair of patents, not only for those connected by a citation. Second, it is symmetric, since similarity has no direction, whereas the citation matrix records directed flows from earlier to later inventions. Third, it is measured from the content of the inventions rather than from the references their applicants chose to make. It therefore captures technological proximity independently of any recorded knowledge flow.

We compute S with two independent methods. The first uses PaECTER (Ghosh et al., 2024), a transformer model for patent text built on BERT for Patents and fine-tuned on patent citations, which maps each document into a 1,024-dimensional vector; the similarity between two patents is the cosine of their vectors. Because raw cosine similarities between transformer embeddings are compressed in a narrow band, we subtract the mean patent vector before computing cosines, which increases the dispersion of pairwise similarities tenfold without affecting their ranking. Since PaECTER is trained to predict citations, however, its notion of similarity is partly informed by the very citation structure we seek to control for. Our second measure therefore uses no citation information. Following Arts et al. (2018, 2023), we represent each patent as a vector of term frequency–inverse document frequency (TF-IDF) weights over the stemmed words of its title, abstract and claims, excluding stopwords, numbers, terms appearing in only one patent and terms appearing in more than 30% of patents. The resulting vocabulary has 21,417 stems, and similarity is the cosine between keyword vectors. Unlike embeddings, this measure has no length limit but does not recognize synonyms.

Figure 4 compares the PaECTER similarity of citing pairs with that of all non-citing pairs. The two distributions are sharply separated: non-citing pairs are centred around zero, whereas most citing pairs lie in the upper tail, with an area under the ROC curve of 0.95. The TF-IDF measure separates the two groups almost as well (area under the curve of 0.94). Citation links

therefore form between technologically close inventions, which confirms empirically the concern of Section 3.3. The two measures rank pairs similarly (Spearman correlation of 0.84) but not identically, so they provide two distinct and partly independent controls.

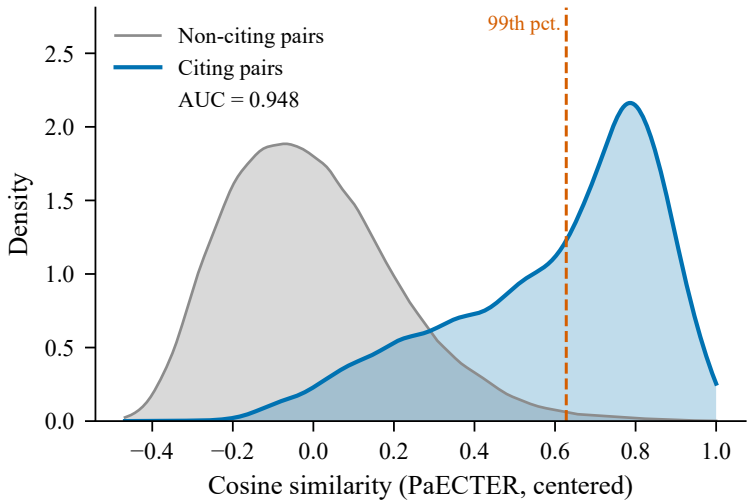


Figure 4: Technological similarity of citing and non-citing pairs

Notes: Density of the PaECTER cosine similarity (centered embeddings of titles, abstracts and claims) for pairs of sample patents linked by a citation and for all other pairs. The dashed line marks the 99th percentile of all pairwise similarities, which defines similar pairs in the partition of Section 5.3. AUC: area under the ROC curve for the classification of citing and non-citing pairs.

5.3 Partition of the knowledge network

Because citation and similarity links strongly overlap, entering the citation matrix and the similarity matrix as two separate networks would produce nearly collinear network terms. Following the logic of Section 3.3, we instead combine them and assign every pair of patents to at most one of three disjoint networks: W_t^b contains pairs that cite each other and are similar, W_t^s pairs that are similar but do not cite each other, and W_t^c pairs that cite each other but are not similar. In our main definition, a pair is similar if its element of S exceeds the 99th percentile of all pairwise similarities, which yields about 33,000 similar pairs. A similarity link becomes active in the period in which the later of the two patents enters the panel, mirroring the activation of citations. All links are binary, and each network is row-normalized separately. Table 3 reports the size of the resulting networks for the two similarity measures.

Table 3: Size of the disjoint networks

Similarity measure	Network	Pairs	Neighbours per obs.
PaECTER	W^b (cited and similar)	7,643	3.23
	W^s (similar, not cited)	25,610	9.04
	W^c (cited, not similar)	5,962	2.27
TF-IDF	W^b (cited and similar)	7,900	3.31
	W^s (similar, not cited)	25,374	9.20
	W^c (cited, not similar)	5,705	2.18

Notes: A pair is similar if its similarity exceeds the 99th percentile of all pairwise similarities. Pairs: unordered patent pairs; all three networks are symmetric, so each pair links both inventions to each other. Neighbours per obs.: average number of active neighbours per patent-year, computed over the 54,341 patent-year observations of the estimation sample.

As a robustness check, we also adopt an alternative, nearest-neighbour definition, under which a pair is similar if either patent is among the 50 most similar patents of the other according to PaECTER. Unlike the threshold definition, which concentrates similar pairs in the densest technology areas, this definition guarantees that every patent has similar neighbours.

The partition gives the three network coefficients of equation (4) a clear interpretation. Co-movement along W_t^s arises between inventions that are technologically close but have no recorded knowledge link, and therefore measures assortativity in its pure form, whereas co-movement along W_t^c arises between inventions that are linked by a citation but are not technologically close, so it cannot be attributed to similarity.

5.4 Internal and external citations

Many citations are internal to the organization that owns the patents, as firms and other institutions build on their own earlier programmes. We classify a citation as *internal* if an applicant of the citing patent is a current owner of the cited patent, matching owners on their Orbis identifiers or, where these are missing, on cleaned names, and as *external* otherwise. Owners are mostly pharmaceutical and biotechnology firms, but also include universities, research institutes and individual inventors, so the distinction refers to the boundaries of the owning organization. Figure 5 shows that the two types of citations are equally close in technology space: 57% of internal and 55% of external links connect patents that are similar under the main definition, and thus belong to W_t^b, while the remainder belong to W_t^c. Owners do not cite technologically closer inventions when they build on their own work than when they build on the work of others. Any difference between the internal and the external spillover estimated in Section 7 can therefore not be attributed to a different degree of technological proximity.

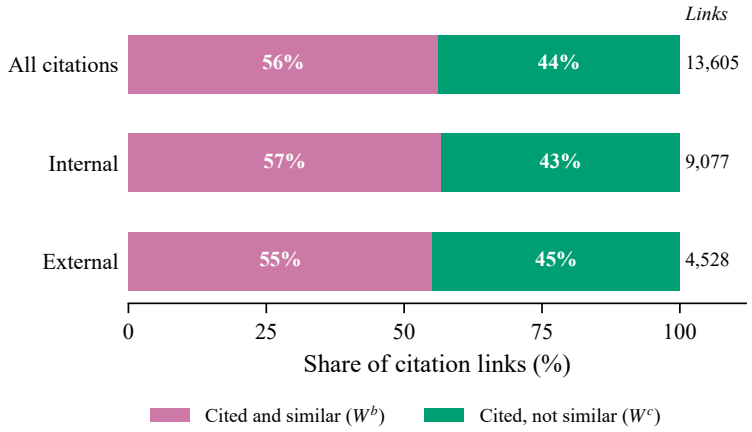


Figure 5: Technological similarity of internal and external citations

Notes: Share of citation links between similar patents (cited and similar, W^b) and between non-similar patents (cited, not similar, W^c), for all citations and separately for internal and external citations; PaECTER similarity, 99th-percentile threshold. Total number of links on the right.

6 Empirical model and methods

6.1 Model specification

The empirical counterpart of the cost equation (1), extended with the partition of Section 5.3, is the following spatial Durbin model:

$$c_{it} = \rho_1 \sum_j w^b_{ijt}\, c_{jt} + \rho_2 \sum_j w^s_{ijt}\, c_{jt} + \rho_3 \sum_j w^c_{ijt}\, c_{jt} + \beta\, x_{it} + \theta \sum_j w_{ijt}\, x_{jt} + Z_{it}\gamma + \alpha_{g(i),t} + \varepsilon_{it}, \quad (7)$$

where c_{it} is the log cumulative cost of patent i in period t, x_{it} its log knowledge exposure defined in equation (6), w_{ijt} the elements of the directed citation matrix W_t, and w^b_{ijt}, w^s_{ijt} and w^c_{ijt} the elements of the three disjoint networks W^b_t, W^s_t and W^c_t. All matrices are binary and row-normalized period by period, so that each network term is the average over a patent's active neighbours in the corresponding network. The peer coefficients ρ_1, ρ_2 and ρ_3 measure the co-movement of costs among inventions that are cited and similar, similar but not cited, and cited but not similar, respectively. The coefficient θ measures the relationship between the cost of an invention and the average exposure of the inventions it cites, that is, the R&D embodied in the knowledge on which its own sources were built. Because exposure is constructed only from cited, and therefore earlier, inventions and is net of shared trials, θ captures a directed flow from upstream R&D to the current project.

The vector Z_{it} collects project-level characteristics that may affect development costs independently of the network. These include, for example, the number of citations a patent makes to other patents in the sample, which measures how embedded the invention is in the network; the number of backward citations to all prior art, which measures the size of the knowledge base on which the invention draws; and the size of the patent family, a standard proxy for the economic importance of an invention.

Finally, $\alpha_{g(i),t}$ denotes fixed effects for each combination of CPC subclass and period, where $g(i)$ is the CPC subclass of patent i. They play a central role in the identification strategy, as development costs vary systematically across technology fields: biologics are typically more expensive to develop than small molecules, and some therapeutic areas require larger and longer trials than others. Costs also evolve over time, as regulatory requirements, trial designs and input prices change. Since Evaluate's estimates rest on a cost-per-patient benchmark that is not specific to the year in which a trial starts, they are expressed at a common, period-average price level, and the fixed effects absorb any residual difference in cost levels across periods. More importantly, technology field and period interact. A technology field may become "hot" in a given period, attracting a wave of investment and raising both the costs of the projects within it and the density of the citations among them, which would generate a spurious association between exposure and cost. Interacting technology field with period absorbs all such shocks, together with common time trends and time-invariant differences across fields. Identification therefore comes only from comparing patents within the same technology field and period, that is, from asking whether, among inventions of the same kind developed at the same time, those that build on costlier prior knowledge are themselves more or less costly. The CPC subclass is the natural level of aggregation for this purpose: it is fine enough to separate the main families of pharmaceutical inventions, yet coarse enough to leave sufficient variation within each cell. Dropping the observations that are the only members of their cell, the estimation sample contains 54,341 patent-year observations.

6.2 Estimation and network impacts

Because the cost of a patent depends on the costs of its neighbours, and vice versa, the peer terms in equation (7) cannot be treated as ordinary regressors: this is the reflection problem (Manski, 1993). Following Castellacci et al. (2026), we address it by estimating equation (7) as a spatial Durbin model with the `nwxtregress` routine (Grieser et al., 2022), a Bayesian estimator of spatial autoregressive models with spatially lagged regressors that accommodates unbalanced panels with time-varying sparse networks.

The routine allows a single endogenous spatial lag of the dependent variable. We assign it to the cite-only network W^c_t, so that ρ_3 is the endogenous parameter that generates the network feedback multiplier $(I-\rho_3 W^c_t)^{-1}$. The lags on W^b_t and W^s_t enter as additional regressors and are estimated jointly with the other coefficients but do not enter the multiplier. The choice reflects the focus of the analysis. The cite-only network is the component of the citation network that cannot be attributed to technological similarity, so it is the one through which a feedback of

costs that is specific to knowledge links can propagate.

As discussed in Section 3.2, the effect of exposure on cost is not given by a single coefficient. In line with that section, we refer to β, the effect of the cost of the cited inventions, as the direct effect of exposure, and to θ, the effect of the cost of the knowledge on which they were built, as the indirect effect. Following LeSage and Pace (2009), the total impact of exposure is the average row sum of the matrix of partial effects, which includes the feedback through the cite-only network and is approximately $(\beta + \theta)/(1 - \rho_3)$, the immediate effect amplified by the network multiplier. The total impact measures the overall elasticity of current cost to the R&D embodied in prior knowledge, and its sign discriminates between the standing-on-shoulders and the crowding hypotheses; the indirect effect isolates the part of this elasticity that originates from the R&D embodied in the sources of the cited inventions, which is the knowledge spillover in the strict sense. We also estimate a pure spatial autoregressive model, obtained by setting $\theta = 0$ in equation (7), to test whether the peer coefficients depend on the inclusion of the exposure channel.

The spatial structure of the model addresses the simultaneity of costs across linked inventions, but not the way the links themselves are formed. Because technologically similar inventions are more likely to cite each other, and are also likely to have similar costs for reasons unrelated to knowledge flows (Section 3.3), part of the citation network consists of links that reflect technological proximity rather than the diffusion of knowledge. The co-movement of costs along citations may therefore capture assortativity rather than spillovers. The CPC subclass $\times$ period fixed effects absorb this assortativity only at the level of broad technology fields. The partition of Section 5.3 addresses it at the level of the individual pair of inventions: by separating the links that arise from technological similarity from those that do not, it controls for the citations that are formed in the network but carry no knowledge spillover, and isolates the citation-specific co-movement (ρ_3) and the exposure spillover (θ) from the co-movement that arises among similar inventions regardless of whether they cite each other (ρ_2). In all the extensions of Section 6.3, the peer terms and the multiplier are those of equation (7), and only the exposure terms are decomposed.

6.3 Decomposing exposure: internal and external knowledge, successes and failures

The exposure terms of equation (7) pool all upstream R&D. We decompose them along the two dimensions introduced in Sections 3.4 and 5.4: whether the upstream knowledge was produced by the same owner, and whether the projects that produced it succeeded or failed.

The two kinds of network terms in equation (7) answer different questions, and each extension acts on the one that matches it. The peer terms relate a patent's cost to the contemporaneous costs of its neighbours. This is where endogenous network formation operates, since similar inventions have similar costs at the same time whether or not knowledge flows between them, and it is therefore the peer term that we decompose by similarity. By contrast, the exposure terms relate a patent's cost to the predetermined R&D embodied in the knowledge it builds on, a directed flow from upstream projects that is the spillover of interest. Which upstream knowledge drives this spillover, whether produced by the same owner or by others and whether from projects that succeeded or failed, is a question about the composition of this flow, and it is therefore the exposure terms that we decompose along these dimensions.

To distinguish internal from external knowledge, we split both exposure terms according to whether the cited invention belongs to the same owner. The own exposure becomes

$$X_{it}^{\text{int}} = \ln\Big(1 + \sum_{j \in \text{int}} \tilde{C}_{j,-i}\Big), \qquad X_{it}^{\text{ext}} = \ln\Big(1 + \sum_{j \in \text{ext}} \tilde{C}_{j,-i}\Big), \tag{8}$$

the log cost of the internally and externally cited inventions, and the upstream exposure becomes

$$x_{it}^{\text{int}} = \frac{1}{n_{it}} \sum_{j \in \text{int}} x_{jt}, \qquad x_{it}^{\text{ext}} = \frac{1}{n_{it}} \sum_{j \in \text{ext}} x_{jt}, \tag{9}$$

where n_{it} is the total number of active cited inventions, so that the two upstream components add up to the row-normalized exposure term $\sum_j w_{ijt} x_{jt}$. The four components replace the two exposure terms of equation (7), with coefficients β^{int}, β^{ext}, θ^{int} and θ^{ext}. The total impact of internal knowledge is approximately $(\beta^{\text{int}} + \theta^{\text{int}})/(1 - \rho_3)$, and that of external knowledge $(\beta^{\text{ext}} + \theta^{\text{ext}})/(1 - \rho_3)$.

To distinguish successes from failures, we sort both exposure terms by the outcome of the cited inventions as known at the time (Section 4.4). For each cited invention j and period t, let s_{jt}^S and s_{jt}^F be the shares of its decided trials known before t to have advanced (successes, S) and to have failed (failures, F), and $s_{jt}^U = 1 - s_{jt}^S - s_{jt}^F$ the share whose outcome is not yet known or that has no decided trial (unknown, U). A patented invention is therefore not classified as a success or a failure as a whole: its cost is attributed to each outcome in proportion to the outcomes of its own trials. The own exposure becomes

$$X_{it}^k = \ln\Big(1 + \sum_{j:\, w_{ijt}>0} \tilde{C}_{j,-i}\, s_{jt}^k\Big), \qquad k \in \{S, F, U\}, \tag{10}$$

the log cost of the cited inventions attributed to outcome k, net of shared trials, and the upstream exposure becomes

$$L_{it}^k = \frac{1}{n_{it}} \sum_{j:\, w_{ijt}>0} x_{jt}\, s_{jt}^k, \qquad k \in \{S, F, U\}, \tag{11}$$

the part of the exposure of the cited inventions that reaches i through inventions of outcome k. Since the shares sum to one, $L_{it}^S + L_{it}^F + L_{it}^U = \sum_j w_{ijt} x_{jt}$: upstream exposure is not altered, only sorted by outcome. Replacing the two exposure terms of equation (7) with their components yields

$$c_{it} = \rho_1 \sum_j w_{ijt}^b\, c_{jt} + \rho_2 \sum_j w_{ijt}^s\, c_{jt} + \rho_3 \sum_j w_{ijt}^c\, c_{jt} + \sum_k \beta_k\, X_{it}^k + \sum_k \theta_k\, L_{it}^k + Z_{it}\gamma + \alpha_{g(i),t} + \varepsilon_{it}. \tag{12}$$

For each outcome, β_k measures the effect of the cost of the cited inventions themselves, one step upstream, and θ_k the effect of the cost of the knowledge on which they were built, two steps upstream. We report both, together with their combined total impact, the sum of the total impacts of X_{it}^k and L_{it}^k, which is approximately $(\beta_k + \theta_k)/(1 - \rho_3)$. The combined impact measures the response of cost to all the R&D embodied in cited inventions of outcome k, and is the counterpart, by outcome, of the total impact of the baseline model. On top of this, its two components show how this response is distributed across the two steps upstream. The components for outcomes not yet known are included in every regression, so that the parts together cover all upstream exposure, but are not reported, since they mix projects still under way with inventions whose trials have no decided outcome.

Finally, we cross the two decompositions, splitting each X_{it}^k and L_{it}^k into an internal and an external part. This allows the effect of a costly success or a costly failure to differ according to whether it was produced by the same owner or by others.

7 Results

7.1 Baseline estimates

Table 4 reports the estimates of equation (7) for the two measures of similarity, each estimated as a Spatial Durbin Model (SDM) and as a pure Spatial AutoRegressive model (SAR). Three results stand out.

Table 4: Baseline estimates

	PaECTER		TF-IDF	
	SDM	SAR	SDM	SAR
ρ_1: cited and similar (W^b)	0.360***	0.359***	0.366***	0.366***
	(0.007)	(0.007)	(0.006)	(0.006)
ρ_2: similar, not cited (W^s)	0.566***	0.576***	0.573***	0.581***
	(0.009)	(0.009)	(0.009)	(0.009)
ρ_3: cited, not similar (W^c)	0.105***	0.099***	0.058***	0.052***
	(0.007)	(0.007)	(0.007)	(0.007)
θ: upstream exposure (W)	0.133***	–	0.141***	–
	(0.012)		(0.012)	
β: own exposure	−0.027***	0.002	−0.030***	0.001
	(0.005)	(0.004)	(0.005)	(0.004)
CPC subclass × period FE	Yes	Yes	Yes	Yes
Observations	54,341	54,341	54,341	54,341
R-squared	0.21	0.21	0.22	0.22

Notes: Dependent variable: log cumulative R&D cost of the patent. Posterior standard deviations in parentheses. All networks binary and row-normalized. All specifications include the number of citations made to sample patents, log backward citations and log family size as controls. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

First, technological similarity is a strong source of cost co-movement. The coefficient on the network of similar inventions that do not cite each other, ρ_2, is about 0.57 under both measures: a patent whose average similar-but-not-cited neighbour is 10% costlier tends to be about 5.5% costlier itself. This co-movement arises without any recorded knowledge link, and therefore captures assortativity in its pure form. The coefficient on inventions that are both cited and similar, ρ_1, is smaller, at about 0.36. Among similar inventions, the presence of a citation is thus not associated with stronger co-movement of costs. Taken together, these estimates confirm the premise of Section 3.3: a large part of the co-movement of costs along citations reflects the technological proximity of the linked inventions rather than the diffusion of knowledge between them. This finding echoes the concern raised by Thompson and Fox-Kean (2005) in the context of the geographic localization of knowledge flows, where much of the apparent concentration of citations disappears once cited and citing patents are matched on finer technological categories, and it is consistent with the strong association between textual similarity and citation documented by Arts et al. (2018).

Second, the citation-specific channel survives the control for similarity. The coefficient on citations between dissimilar inventions, ρ_3, is positive and significant at the 1% level under both measures, at 0.105 with PaECTER and 0.058 with TF-IDF, which implies that a patent whose average cited-but-dissimilar neighbour is 10% costlier tends to be about 0.6–1.0% costlier itself. Because these links connect inventions that are not technologically close, this co-movement cannot be attributed to assortativity, and, more importantly, the rejection of the pure-assortativity hypothesis ($\rho_3 = 0$) does not depend on the measure. In this respect, our results support the view that patent citations carry information on knowledge flows beyond the technological proximity of the linked inventions (Jaffe et al., 1993), while confirming that the noise introduced by citations unrelated to actual knowledge use (Alcácer and Gittelman, 2006) is a relevant concern in this setting.

Third, upstream R&D is associated with higher, not lower, development costs. The coefficient on upstream exposure, θ, is 0.133 with PaECTER and 0.141 with TF-IDF, both significant at the 1% level: a 10% higher R&D investment in the knowledge on which a patent's sources were

built goes along with a 1.3–1.4% higher cost of the patent itself. The coefficient on own exposure, β, is instead small and negative in the spatial Durbin model, at -0.027 and -0.030, whereas it is close to zero in the pure spatial autoregressive model. The difference reflects the correlation between the two exposure terms; that is, own exposure and upstream exposure are built from the same cited inventions, so that when upstream exposure is omitted, own exposure partly absorbs its positive association with cost. Conditional on the R&D embodied in the knowledge behind them, building on costlier cited inventions is associated with slightly lower development costs. The positive association between upstream R&D and cost contrasts with evidence that spillovers from rivals' research raise the productivity of drug discovery (Henderson and Cockburn, 1996), and suggests that the input side of pharmaceutical R&D behaves differently from the output one.

Finally, the peer coefficients of the spatial Durbin and pure spatial autoregressive models are very close: including the exposure terms moves ρ_2 by about 0.01 and ρ_3 by about 0.006, well within two posterior standard deviations. The co-movement of costs among linked inventions and the spillover of upstream exposure are thus distinct channels, and including the latter does not materially alter the former.

7.2 The impact of upstream R&D

Table 5: Impacts of exposure on cost

	PaECTER	TF-IDF
Direct: own exposure, β	-0.027***	-0.030***
	(0.005)	(0.005)
Indirect: upstream exposure, θ	0.133***	0.141***
	(0.012)	(0.012)
Total, $\dfrac{\beta+\theta}{1-\rho_3}$	0.118***	0.117***
	(0.011)	(0.011)
Multiplier, $\dfrac{1}{1-\rho_3}$	1.117	1.062

Notes: Computed from the spatial Durbin estimates of Table 4. Direct: coefficient on the cost of the cited inventions (own exposure); indirect: coefficient on the cost of the knowledge on which they were built (upstream exposure). Total: average total impact of exposure on log cumulative cost, including the feedback through the cite-only network; with row-normalized networks it is approximately the immediate effect $\beta+\theta$ amplified by the multiplier. The posterior standard deviation of the total is that of $\beta+\theta$ scaled by the multiplier. Posterior standard deviations in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.10$.

Table 5 reports the direct, indirect and total effects of exposure on cost. The total impact is 0.118 with PaECTER and 0.117 with TF-IDF, both significant at the 1% level: a 10% increase in the R&D embodied in the knowledge an invention builds on is associated with a 1.1–1.2% increase in the cost of its development. The sign of the total impact discriminates between the two hypotheses of Section 3.1. It is positive, which supports the crowding hypothesis over the standing-on-shoulders hypothesis: on average, building on costlier prior knowledge is associated with more expensive development. This result is consistent with the burden-of-knowledge argument, according to which the accumulation of prior knowledge raises the effort required of later innovators (Jones, 2009), and with evidence that new ideas become progressively harder to find (Bloom et al., 2020). In pharmaceuticals, it resonates with the observation that

crowded technological areas are characterized by lower returns to R&D (Pammolli et al., 2011), and with the argument that each new drug must demonstrate an advantage over an expanding set of existing therapies, which raises the size and cost of the trials required for approval (Scannell et al., 2012). By contrast, the cumulative-innovation view, in which later innovators benefit from building on the work of earlier ones (Scotchmer, 1991), does not appear to dominate on the cost side.

The decomposition of the total impact is equally informative. The positive total impact is driven by the indirect effect: it originates from the R&D embodied in the knowledge on which a patent's sources were built (θ of 0.13–0.14), whereas the direct effect of the cost of the cited inventions themselves is small and negative. The cost-raising effect of prior R&D is therefore a genuinely network phenomenon, transmitted through the chain of citations rather than confined to the direct relationship between an invention and the prior art it cites. The feedback through the cite-only network raises the total impact by 6–12% over the immediate effect $\beta+\theta$, consistent with the multiplier of Section 3.2. The predominance of the indirect component is in line with the evidence that knowledge flows propagate through innovation networks well beyond direct links (Acemoglu et al., 2016; Liu and Ma, 2021), and complements the within-firm accounting of the cost of drug development (DiMasi et al., 2016), in which the costs of related projects enter only when they are borne by the same developer.

7.3 Robustness

Table 6 examines the sensitivity of the main results to two alternative choices, using PaECTER as the measure of similarity. Column (2) replaces the threshold definition of similar pairs with the nearest-neighbour definition, under which a pair is similar if either patent is among the 50 most similar patents of the other. Column (3) uses non-normalized matrices, in which network terms, including upstream exposure, are sums rather than averages over neighbours.

Table 6: Robustness

	(1) Baseline	(2) Nearest neighbours	(3) Non-normalized
ρ_1: cited and similar	0.360***	0.473***	0.042***
	(0.007)	(0.006)	(0.001)
ρ_2: similar, not cited	0.566***	0.543***	0.025***
	(0.009)	(0.017)	(0.001)
ρ_3: cited, not similar	0.105***	0.076***	0.024***
	(0.007)	(0.008)	(0.001)
θ: upstream exposure	0.133***	0.169***	0.003**
	(0.012)	(0.012)	(0.001)
Total impact of exposure	0.118***	0.116***	0.026***
	(0.011)	(0.011)	(0.005)
Observations	54,341	54,341	54,341
R-squared	0.21	0.16	0.12

Notes: Spatial Durbin model, PaECTER similarity. (1): baseline of Table 4. (2): a pair is similar if either patent is among the 50 most similar patents of the other. (3): binary, non-normalized matrices and upstream exposure summed over the cited inventions; coefficients are per link. Total impact of exposure as in Table 5. Controls and CPC subclass × period fixed effects as in Table 4. Posterior standard deviations in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

The results are robust along both dimensions. Under the nearest-neighbour definition, which guarantees similar neighbours to every patent and yields a similarity network about 2.7 times denser, ρ_3 remains positive and significant, ρ_2 remains the largest peer coefficient, and the total impact of exposure is 0.116, almost identical to the baseline. With non-normalized matrices, the coefficients measure the effect of a single neighbour rather than of the average one, so their magnitudes are not comparable with the baseline; their signs and significance, however, are unchanged, and the total impact of exposure remains positive and significant.

7.4 Internal and external knowledge

As lots of citations are internal to the owner of the patents (Section 5.4), the cost effect estimated in Table 5 may reflect owners building on their own earlier programmes rather than the diffusion of knowledge across organizational boundaries. Table 7 reports the estimates with both exposure terms split into their internal and external components, as in equations (8) and (9), together with the total impact of each.

Table 7: Internal and external exposure

	PaECTER	TF-IDF
Coefficients		
β^{int}: own exposure, internal	−0.014*	−0.021***
	(0.008)	(0.008)
β^{ext}: own exposure, external	−0.036***	−0.036***
	(0.006)	(0.006)
θ^{int}: upstream exposure, internal	0.158***	0.164***
	(0.018)	(0.018)
θ^{ext}: upstream exposure, external	0.103***	0.116***
	(0.015)	(0.015)
Total impacts		
Internal, $\dfrac{\beta^{\text{int}} + \theta^{\text{int}}}{1 - \rho_3}$	0.162***	0.152***
	(0.016)	(0.015)
External, $\dfrac{\beta^{\text{ext}} + \theta^{\text{ext}}}{1 - \rho_3}$	0.075***	0.085***
	(0.015)	(0.015)
Test internal = external (p-value)	<0.001	0.001
Observations	54,341	54,341
R-squared	0.21	0.22

Notes: Spatial Durbin model with own and upstream exposure split by owner, as in equations (8) and (9). Total impacts include the feedback through the cite-only network; their posterior standard deviations are those of $\beta + \theta$ scaled by the multiplier. Peer networks, controls and CPC subclass × period fixed effects as in Table 4. Posterior standard deviations in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Both total impacts are positive and highly significant (Figure 6). A 10% increase in the R&D embodied in internal knowledge is associated with a 1.5–1.6% increase in development cost, and a 10% increase in the R&D embodied in external knowledge with a 0.8% increase. The difference between the two is significant under both measures. The pattern of the components mirrors that of the baseline: in both cases the positive effect comes from upstream exposure, while the direct effect of the cost of the cited inventions is small and negative, and more so for external knowledge. The effect is therefore about twice as strong within the owning organization,

consistent with the view that owners building on their own costly programmes carry forward their cost structures, but it is not confined to it. Since internal and external citations are equally close in technology space (Figure 5), the gap between the two cannot be attributed to a different degree of technological proximity. The positive association between prior R&D and development costs thus extends across organizational boundaries, which is the externality that the within-firm accounting of drug development costs does not capture (DiMasi et al., 2016).

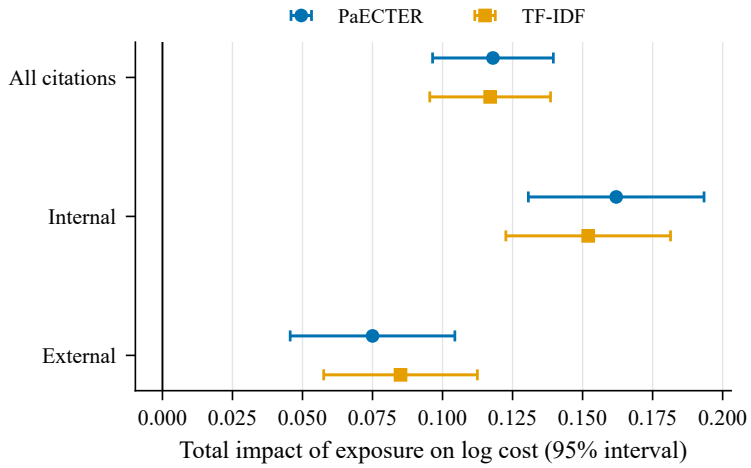


Figure 6: Total impact of exposure on cost, internal and external knowledge

Notes: Total impacts of exposure on log cumulative cost with 95% intervals for the two similarity measures. All citations (Table 5); internal and external (Table 7).

7.5 Successes and failures

Table 8 reports the total impacts of exposure to cited inventions of each outcome, estimated from equation (12), first pooled across owners and then separately for internal and external knowledge. Table 9 decomposes each total impact into the part due to the cost of the cited inventions themselves (β_k) and the part due to the cost of the knowledge on which they were built (θ_k).

The cost spillover documented above comes from building on successful projects. A 10% increase in the R&D embodied in cited successes is associated with a 1.3–1.5% higher development cost, network feedback included, whereas cited failures, pooled across owners, show no significant association with cost. The difference between the two is significant with PaECTER ($p = 0.002$) and marginally so with TF-IDF ($p = 0.063$).

The split by owner shows that the pooled zero for failures hides two opposite patterns (Panel B). Within the owning organization, costly failures are associated with even higher later costs than successes, with total impacts of 0.35–0.37 against 0.22–0.26. Across organizations, by contrast, the costly failures of others are associated with lower costs of later development, with a total impact of -0.145 with PaECTER ($p < 0.01$) and -0.085 with TF-IDF ($p < 0.10$), whereas the successes of others are only weakly associated with higher costs (0.05, $p < 0.10$). Owners thus appear to repeat the costly paths of their own unsuccessful programmes, while they learn from the failures of others and avoid the paths that proved unpromising, a pattern consistent with the evidence that firms respond to competitors' failures by redirecting their own development (Krieger, 2021).

The decomposition in Table 9 shows that the two steps upstream pull in opposite directions. For successes, the cost of the cited projects themselves is associated with lower development cost (-0.05 to -0.06), and the cost of the knowledge behind them with higher cost (0.18–0.21). Building directly on costly successes thus appears mildly cost-saving, as validated targets, doses and endpoints can be reused, whereas the deep and costly knowledge base behind them marks

Table 8: Impact of exposure by outcome of the cited inventions

	PaECTER	TF-IDF
Panel A: pooled across owners		
Cited successes	0.150***	0.134***
	(0.019)	(0.018)
Cited failures	−0.009	0.044
	(0.043)	(0.040)
Test successes = failures (p-value)	0.002	0.063
Panel B: by owner		
Internal successes	0.257***	0.222***
	(0.028)	(0.027)
Internal failures	0.352***	0.365***
	(0.074)	(0.070)
External successes	0.047*	0.045*
	(0.027)	(0.025)
External failures	−0.145***	−0.085*
	(0.053)	(0.050)
Observations	54,341	54,341

Notes: Total impact of own and upstream exposure to cited inventions of each outcome, $(\beta_k + \theta_k)/(1 - \rho_3)$, from the spatial Durbin model of equation (12). Successes (failures): cited inventions whose trials were known to have advanced (failed) before the citing period. Components for outcomes not yet known are included but not reported. Peer networks, controls and CPC subclass × period fixed effects as in Table 4. Posterior standard deviations in parentheses, obtained from the standard deviation of $\beta_k + \theta_k$ scaled by the network multiplier. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

a mature, crowded field in which development is more expensive. For failures the pattern is reversed: the cost of the cited failures themselves is associated with higher development cost (0.07–0.08), and the knowledge behind them with lower cost (−0.09 with PaECTER, not significant with TF-IDF). The pattern holds both for internal and for external knowledge, with the exception of internal failures, for which both components are positive.

8 Discussion

Our results point to four main findings. First, although technological similarity accounts for most of the co-movement of costs among related inventions, a smaller but substantial portion of this co-movement is driven by a citation-specific channel that survives the control for similarity. Second, exposure to costlier prior R&D is associated with higher, rather than lower, development costs: a 10% increase in the R&D embodied in the knowledge an invention builds on goes along with a cost about 1.2% higher, an association driven by the knowledge on which the cited inventions were themselves built. Third, the association is about twice as strong for knowledge produced within the owning organization as for knowledge produced by others, but it remains positive and significant across organizational boundaries. Fourth, it originates from upstream projects that succeeded: costly failures are associated with higher later costs when they belong to the same owner, and with lower later costs when they belong to others.

The prevalence of crowding over standing-on-shoulders effects can be understood in light of the institutional features of pharmaceutical development. Prior research generates knowledge

Table 9: Decomposition into own and upstream exposure

	PaECTER		TF-IDF	
	Own	Upstream	Own	Upstream
Panel A: pooled across owners				
Cited successes	−0.060***	0.210***	−0.050***	0.184***
	(0.007)	(0.021)	(0.007)	(0.019)
Cited failures	0.084***	−0.094**	0.065***	−0.021
	(0.011)	(0.045)	(0.010)	(0.042)
Panel B: by owner				
Internal successes	−0.029***	0.285***	−0.022**	0.243***
	(0.011)	(0.032)	(0.010)	(0.031)
Internal failures	0.111***	0.241***	0.090***	0.275***
	(0.023)	(0.076)	(0.022)	(0.071)
External successes	−0.091***	0.138***	−0.079***	0.124***
	(0.009)	(0.027)	(0.009)	(0.026)
External failures	0.106***	−0.251***	0.081***	−0.166***
	(0.013)	(0.054)	(0.012)	(0.051)

Notes: Total impacts of the own-exposure component X_{it}^k (cost of the cited inventions themselves) and of the upstream-exposure component L_{it}^k (cost of the knowledge on which they were built), from the same estimates as Table 8. For each outcome, the two entries sum to the corresponding entry of Table 8. Posterior standard deviations in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

that later projects can exploit, but it also raises the bar that they must clear. A new compound entering a field in which substantial investment has already been made must demonstrate an advantage over a growing set of existing therapies, which requires larger and longer trials (Scannell et al., 2012), and it competes with parallel efforts for the same patients, investigators and trial sites (Pammolli et al., 2011). Projects that build on a costly body of prior work are also likely to be follow-on inventions in mature segments, where the remaining opportunities are harder to achieve (Jones, 2009; Bloom et al., 2020). The outcome results sharpen this interpretation. The cost premium is concentrated in knowledge from successful programmes, precisely those that establish an effective therapy and define the standard that later entrants must beat, while the R&D embodied in the cited successes themselves is associated with slightly lower costs, consistent with the reuse of validated targets, doses and endpoints. Intellectual property provides an additional reason why building on costly and successful prior knowledge may be associated with higher costs. Upstream inventions that embody substantial investment are typically protected by patents, so that follow-on projects must obtain licences, invent around existing claims or bear the risk of infringement. Overlapping rights can give rise to patent thickets and to an anticommons in biomedical research, in which the cost of assembling the rights required for development rises with the density of prior protection (Heller and Eisenberg, 1998; Shapiro, 2001; Cockburn et al., 2010), and patents have been shown to reduce follow-on innovation in some technology fields (Williams, 2013; Galasso and Schankerman, 2015), although not uniformly (Sampat and Williams, 2019).

The split by owner and by outcome points to different mechanisms within and across organizations. Within the owning organization, the larger association is consistent with owners carrying forward the cost structure of their own programmes, such as trial designs, comparators and target indications, and the positive association with own failures suggests that this persistence extends to lines of research that did not succeed. Across organizations, the smaller but significant association shares the logic of the product-market rivalry effect identified by Bloom

et al. (2013), that is, the R&D of others is a source of knowledge but also of competition and of legal constraints, and on the cost side these forces dominate. The costly failures of others, by contrast, are associated with lower later costs, in line with the evidence that firms learn from the failures of competitors and abandon unpromising paths (Krieger, 2021; Magazzini et al., 2012; Chiou et al., 2016). Knowledge spillovers on the cost side thus operate in both directions, and which direction prevails depends on whether the upstream knowledge comes from success or failure and from the same or another organization.

Controlling for assortativity is essential to these conclusions. Technologically similar inventions share therapeutic areas, molecules and regulatory pathways, and therefore have similar costs regardless of any knowledge flow; without a control for similarity, the co-movement of costs along citations would mix diffusion with technological proximity. The central result of the paper is that the cost spillover survives this control: both the co-movement along citations between dissimilar inventions and the association between upstream exposure and cost remain positive and significant once similarity is accounted for, under two independent measures of similarity and two definitions of similar pairs.

Some caveats apply. First, our estimates are conditional associations rather than causal effects: although exposure is predetermined and the partition controls for assortativity on the observable content of inventions, unobserved determinants of both link formation and cost may remain. Second, the peer terms on the similarity-based networks are treated as exogenous, even though they involve contemporaneous costs. Finally, the sample is restricted to trials linked to patents and with a cost estimate, and programmes that advanced are over-represented among them, particularly at the transition from Phase 2 to Phase 3. The failures we observe may therefore differ from the typical failed programme, and the effects of upstream failures are estimated less precisely than those of successes.

9 Conclusions

This paper asks whether R&D expenditures generate cost spillovers across technologically related pharmaceutical projects. Linking trial-level costs to a citation network of U.S. patents, and measuring technological similarity from the full text of their claims, we estimate spatial econometric models that separate the co-movement of costs due to technological proximity from the co-movement that travels along citation links, and we sort the knowledge each invention builds on by its owner and by the outcome of the projects that produced it. Four findings emerge. Technological similarity accounts for most of the co-movement of costs among related inventions, so that controlling for it is necessary to avoid overstating spillovers. Once similarity is controlled for, a citation-specific channel remains, and exposure to costlier prior R&D is associated with higher, not lower, development costs: a 10% increase in the R&D embodied in the knowledge an invention builds on goes along with a 1.2% higher cost, driven by the knowledge on which the cited inventions were themselves built. The association is about twice as strong for knowledge produced within the owning organization, but it extends across organizational boundaries. Finally, it originates from upstream programmes that succeeded, whereas the costly failures of other organizations are associated with lower costs of later development.

These results complement the within-firm accounting of the cost of drug development with an across-project dimension. The cost of a new drug depends not only on the failures borne by its developer, but also on the costly knowledge accumulated by others in the same technological area. In pharmaceuticals, crowding, rising evidentiary standards and the protection of prior inventions appear to outweigh the cost savings that cumulative knowledge could offer, at least when that knowledge comes from successful programmes. The opposite association for the failures of others suggests that information on what did not work has value for later developers, and that the public disclosure of trial results, including negative ones, through registries such as ClinicalTrials.gov may help contain the cost of subsequent development.

Several extensions are left for future work. First, a quasi-experimental source of variation in upstream exposure, such as unexpected outcomes of upstream programmes, would allow a causal interpretation of the associations documented here. Second, the cost premium associated with building on costly prior knowledge may be compensated by a higher probability that later projects advance through development: relating exposure to the success of the citing projects themselves would show whether the additional cost buys a higher chance of reaching the market.

Acknowledgements

Data

Data on trial–patent links are from the PubMed Knowledge Graph 2.0, trial registrations from ClinicalTrials.gov and trial outcome labels from the Clinical Trial Outcome benchmark, all publicly available. Data on trial costs and programme histories are from Evaluate Pharma®, and patent data are from Orbis Intellectual Property (Bureau van Dijk, a Moody's Analytics company); both are commercial databases available upon subscription.

Research support and disclosures

Michele Liberatore acknowledges financial support from a doctoral scholarship of the PhD programme in Social Sciences for Sustainability and Wellbeing (S3W), IMT School for Advanced Studies Lucca and University of Florence, funded by the European Union – NextGenerationEU under the Italian National Recovery and Resilience Plan (NRRP), Mission 4, Component 1, Investment 4.1, M.D. 629/2024 (CUP B12B24000570007). The authors declare that they have no financial relationships or other potential conflicts of interest relevant to this work.

Use of Generative Artificial Intelligence

During the preparation of this work, the authors used Claude for grammar, spelling and clarity edits, and for debugging analysis code. After using this tool, the authors reviewed, verified and edited all content and code as needed and take full responsibility for the content of the article.

References

Acemoglu, D., Akcigit, U., and Kerr, W. R. (2016). Innovation network. *Proceedings of the National Academy of Sciences*, 113(41):11483–11488.

Alcácer, J. and Gittelman, M. (2006). Patent citations as a measure of knowledge flows: The influence of examiner citations. *The Review of Economics and Statistics*, 88(4):774–779.

Arqué-Castells, P. and Spulber, D. F. (2022). Measuring the private and social returns to R&D: Unintended spillovers versus technology markets. *Journal of Political Economy*, 130(7):1860–1918.

Arts, S., Cassiman, B., and Gomez, J. C. (2018). Text matching to measure patent similarity. *Strategic Management Journal*, 39(1):62–84.

Arts, S., Cassiman, B., and Hou, J. (2023). Position and differentiation of firms in technology space. *Management Science*, 69(12):7253–7265.

Arts, S., Hou, J., and Gomez, J. C. (2021). Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures. *Research Policy*, 50(2):104144.

Azoulay, P., Graff Zivin, J. S., Li, D., and Sampat, B. N. (2019). Public R&D investments and private-sector patenting: Evidence from NIH funding rules. *The Review of Economic Studies*, 86(1):117–152.

Bloom, N., Jones, C. I., Van Reenen, J., and Webb, M. (2020). Are ideas getting harder to find? *American Economic Review*, 110(4):1104–1144.

Bloom, N., Schankerman, M., and Van Reenen, J. (2013). Identifying technology spillovers and product market rivalry. *Econometrica*, 81(4):1347–1393.

Bramoullé, Y., Djebbari, H., and Fortin, B. (2009). Identification of peer effects through social networks. *Journal of Econometrics*, 150(1):41–55.

Budish, E., Roin, B. N., and Williams, H. (2015). Do firms underinvest in long-term research? Evidence from cancer clinical trials. *American Economic Review*, 105(7):2044–2085.

Castellacci, F., Raknerud, A., Riccaboni, M., and Rybalka, M. (2026). Publicly funded R&D networks, knowledge spillovers and innovation performance: Evidence from Norway. *European Economic Review*, 190:105469.

Chiou, J.-Y., Magazzini, L., Pammolli, F., and Riccaboni, M. (2016). Learning from successes and failures in pharmaceutical R&D. *Journal of Evolutionary Economics*, 26(2):271–290.

Cockburn, I. M., MacGarvie, M. J., and Müller, E. (2010). Patent thickets, licensing and innovative performance. *Industrial and Corporate Change*, 19(3):899–925.

DiMasi, J. A., Grabowski, H. G., and Hansen, R. W. (2016). Innovation in the pharmaceutical industry: New estimates of R&D costs. *Journal of Health Economics*, 47:20–33.

DiMasi, J. A., Hansen, R. W., and Grabowski, H. G. (2003). The price of innovation: New estimates of drug development costs. *Journal of Health Economics*, 22(2):151–185.

DiMasi, J. A., Hansen, R. W., Grabowski, H. G., and Lasagna, L. (1991). Cost of innovation in the pharmaceutical industry. *Journal of Health Economics*, 10(2):107–142.

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

Gao, C., Pradeepkumar, J., Das, T., Thati, S., and Sun, J. (2024). Automatically labeling clinical trial outcomes: A large-scale benchmark for drug development. arXiv preprint 2406.10292.

Ghosh, M., Erhardt, S., Rose, M. E., Buunk, E., and Harhoff, D. (2024). PaECTER: Patent-level representation learning using citation-informed transformers. arXiv preprint 2402.19411.

Goldsmith-Pinkham, P. and Imbens, G. W. (2013). Social networks and the identification of peer effects. *Journal of Business & Economic Statistics*, 31(3):253–264.

Grieser, W., LeSage, J., and Zekhnini, M. (2022). Industry networks and the geography of firm behavior. *Management Science*, 68(8):6163–6183.

Griliches, Z. (1979). Issues in assessing the contribution of research and development to productivity growth. *The Bell Journal of Economics*, 10(1):92–116.

Heller, M. A. and Eisenberg, R. S. (1998). Can patents deter innovation? The anticommons in biomedical research. *Science*, 280(5364):698–701.

Henderson, R. and Cockburn, I. (1996). Scale, scope, and spillovers: The determinants of research productivity in drug discovery. *The RAND Journal of Economics*, 27(1):32–59.

Hsieh, C.-S., König, M. D., and Liu, X. (2025). Endogenous technology spillovers in R&D collaboration networks. *The RAND Journal of Economics*, 56(4):419–443.

Jaffe, A. B. (1986). Technological opportunity and spillovers of R&D: Evidence from firms' patents, profits, and market value. *American Economic Review*, 76(5):984–1001.

Jaffe, A. B., Trajtenberg, M., and Henderson, R. (1993). Geographic localization of knowledge spillovers as evidenced by patent citations. *The Quarterly Journal of Economics*, 108(3):577–598.

Johnsson, I. and Moon, H. R. (2021). Estimation of peer effects in endogenous social networks: Control function approach. *The Review of Economics and Statistics*, 103(2):328–345.

Jones, B. F. (2009). The burden of knowledge and the "death of the Renaissance man": Is innovation getting harder? *The Review of Economic Studies*, 76(1):283–317.

Kelly, B., Papanikolaou, D., Seru, A., and Taddy, M. (2021). Measuring technological innovation over the long run. *American Economic Review: Insights*, 3(3):303–320.

König, M. D., Liu, X., and Zenou, Y. (2019). R&D networks: Theory, empirics, and policy implications. *The Review of Economics and Statistics*, 101(3):476–491.

Krieger, J., Li, D., and Papanikolaou, D. (2022). Missing novelty in drug development. *The Review of Financial Studies*, 35(2):636–679.

Krieger, J. L. (2021). Trials and terminations: Learning from competitors' R&D failures. *Management Science*, 67(9):5525–5548.

Kuhn, J., Younge, K., and Marco, A. (2020). Patent citations reexamined. *The RAND Journal of Economics*, 51(1):109–132.

LeSage, J. and Pace, R. K. (2009). *Introduction to Spatial Econometrics*. Chapman and Hall/CRC, Boca Raton.

Liu, E. and Ma, S. (2021). Innovation networks and R&D allocation. NBER Working Paper 29607, National Bureau of Economic Research.

Lucking, B., Bloom, N., and Van Reenen, J. (2019). Have R&D spillovers declined in the 21st century? *Fiscal Studies*, 40(4):561–590.

Magazzini, L., Pammolli, F., and Riccaboni, M. (2012). Learning from failures or failing to learn? Lessons from pharmaceutical R&D. *European Management Review*, 9(1):45–58.

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies*, 60(3):531–542.

Myers, K. R. and Lanahan, L. (2022). Estimating spillovers from publicly funded R&D: Evidence from the US Department of Energy. *American Economic Review*, 112(7):2393–2423.

Orsenigo, L., Pammolli, F., and Riccaboni, M. (2001). Technological change and network dynamics: Lessons from the pharmaceutical industry. *Research Policy*, 30(3):485–508.

Pammolli, F., Magazzini, L., and Riccaboni, M. (2011). The productivity crisis in pharmaceutical R&D. *Nature Reviews Drug Discovery*, 10(6):428–438.

Sampat, B. and Williams, H. L. (2019). How do patents affect follow-on innovation? Evidence from the human genome. *American Economic Review*, 109(1):203–236.

Scannell, J. W., Blanckley, A., Boldon, H., and Warrington, B. (2012). Diagnosing the decline in pharmaceutical R&D efficiency. *Nature Reviews Drug Discovery*, 11(3):191–200.

Scotchmer, S. (1991). Standing on the shoulders of giants: Cumulative research and the patent law. *Journal of Economic Perspectives*, 5(1):29–41.

Sertkaya, A., Beleche, T., Jessup, A., and Sommers, B. D. (2024). Costs of drug development and research and development intensity in the US, 2000–2018. *JAMA Network Open*, 7(6):e2415445.

Shapiro, C. (2001). Navigating the patent thicket: Cross licenses, patent pools, and standard setting. *Innovation Policy and the Economy*, 1:119–150.

Thompson, P. and Fox-Kean, M. (2005). Patent citations and the geography of knowledge spillovers: A reassessment. *American Economic Review*, 95(1):450–460.

Williams, H. L. (2013). Intellectual property rights and innovation: Evidence from the human genome. *Journal of Political Economy*, 121(1):1–27.

Wouters, O. J., McKee, M., and Luyten, J. (2020). Estimated research and development investment needed to bring a new medicine to market, 2009–2018. *JAMA*, 323(9):844–853.

Xu, J., Yu, C., Xu, J., Torvik, V. I., Kang, J., Sung, M., Song, M., Bu, Y., and Ding, Y. (2025). PubMed knowledge graph 2.0: Connecting papers, patents, and clinical trials in biomedical science. *Scientific Data*.

Zacchia, P. (2020). Knowledge spillovers through networks of scientists. *The Review of Economic Studies*, 87(4):1989–2018.

Zarin, D. A., Tse, T., Williams, R. J., Califf, R. M., and Ide, N. C. (2011). The ClinicalTrials.gov results database: Update and key issues. *New England Journal of Medicine*, 364(9):852–860.