NBER WORKING PAPER SERIES

REVISITING THE ENTREPRENEURIAL COMMERCIALIZATION OF ACADEMIC SCIENCE: EVIDENCE FROM "TWIN" DISCOVERIES

Matt Marx
David H. Hsu

Working Paper 28203
http://www.nber.org/papers/w28203

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 2020

We are grateful for feedback from Christine Beckman, Kristoph Kleiner, and participants of the Wharton Technology & Innovation Conference; the Global Entrepreneurship & Innovation Conference; the Strategy Science Conference, West Coast Research Symposium at Stanford; the Crete Workshop on Innovation & Creativity; as well as department seminars at Washington University in St. Louis, Cornell University, and KAIST. We thank Chris Ackerman, Rafael Castro, Andrea Contigiani, and Luming Yang for excellent research assistance. We also thank Guan-Cheng Li for data on patent-to-paper citations, Michael Ewens for his USPTO/ VentureSource concordance, and Kyle Myers for his USPTO/Crunchbase/SBIR concordance. We acknowledge research support from Boston University and the Mack Center for Innovation Management at the University of Pennsylvania. This work was supported by National Science Foundation grant 1360228. Errors and omissions are ours The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

Revisiting the Entrepreneurial Commercialization of Academic Science: Evidence from "Twin" Discoveries

Matt Marx and David H. Hsu

NBER Working Paper No. 28203
December 2020
JEL No. L26

ABSTRACT

Which factors shape the commercialization of academic scientific discoveries via startup formation? Prior literature has identified several contributing factors but does not address the fundamental problem that the commercial potential of a nascent discovery is generally unobserved, which potentially confounds inference. We construct a sample of approximately 20,000 "twin" scientific articles, which allows us to hold constant differences in the nature of the advance and more precisely examine characteristics that predict startup commercialization. In this framework, several commonly-accepted factors appear not to influence commercialization. However, we find that teams of academic scientists whose former collaborators include "star" serial entrepreneurs are much more likely to commercialize their own discoveries via startups, as are more interdisciplinary teams of scientists.

Matt Marx
Questrom School of Business
Boston University
595 Commonwealth Avenue
Boston, MA 02215
and NBER
mattmarx@bu.edu

David H. Hsu
University of Pennsylvania
Wharton School
2000 Steinberg-Dietrich Hall
Philadelphia, PA 19104
dhsu@wharton.upenn.edu

1 Introduction & Motivation

The technologies underlying many successful companies including Google's PageRank search algorithm, E-Ink's electronic paper, RSA's cryptography algorithm, and Genentech's recombinant growth hormone were discovered by scientists at universities who then commercialized via startups. Consider Amnon Shashua, professor of computer science at Hebrew University, who published many articles applying computer vision to traffic safety, including "Forward Collision Warning with a Single Camera." Shashua might have left his work in the public domain for others to possibly exploit but instead self-commercialized his research by co-founding Mobileye, which supplies the driver-assistance systems in many vehicles and became Israel's largest startup acquisition when sold to Intel for $15.3B in 2017.

With universities increasingly concerned with economic development alongside their longstanding teaching and research missions, scholars have sought to better understand the factors that explain academic entrepreneurship. Rothaermel, Agung & Jiang (2007) and Markman, Siegel & Wright (2008) catalog 175+ such papers. Commercialization can take several forms including technology licensing to established firms—and we do not claim that a startup is always the optimal vehicle for commercialization—but the literature has highlighted several reasons to understand new venture formation from academia. First, technologies developed in university labs are typically more embryonic than their industrial lab counterparts (Jensen and Thursby 2001). As a result, absent new venture development these discoveries may go uncommercialized (Hsu and Bernstein 1997). Second, commercializing discoveries via venture formation addresses the changing nature of careers in academic science. STEM doctoral degree awardees in the U.S. have long exceeded the number of available academic jobs (Cyranoski et al. 2011) while there has been a steady increase in university spinoffs. Because graduate students are instrumental in academic entrepreneurship (Hayter, Lubynsky, and Maroulis 2017), there are scientific labor market implications of commercializing science. Finally, from an economic development standpoint, startups are disproportionately involved in job growth (Haltiwanger, Jarmin, and Miranda 2013), and academic ventures tend to locate near prominent research scientists (Zucker, Darby, and Brewer 1998), so regional growth may be spurred by venture formation.

Work on commercializing academic research via (discovering scientist-involved) startup formation has focused on two sets of antecedents. A first view, which we call the *resource munificence* perspective, claims that entrepreneurial opportunities proceed to commercialization based on resources available, often within a given geography. Resources could include financial capital (Samila and Sorenson 2011) and know-how, spanning technical as well as commercial domains, and may take place at various levels of analysis including groups (e.g., workshops) and even institutions (e.g., university collaborative relations with private enterprises). Financial capital for developing entrepreneurial opportunities such as from venture capitalists (VCs) is thought to be particularly sensitive to geographic co-location. Resources flow based on researcher or institutional prestige, so this literature has also examined the role of status (Stuart, Hoang, and Hybels 1999).

A second perspective, which we term the *discovery team composition* view, highlights the configuration and social context of the team that discovers the scientific advance. This branch of literature suggests that scientific teams with exposure to peers who have experience in commercializing science can

substantially increase the propensity of engaging in entrepreneurship due to awareness, demonstration effects, professional legitimization, and experience with commercialization (e.g., Stuart & Ding (2006)). Team composition itself can also impact entrepreneurial opportunity recognition and commercialization outcomes such as through social networks and experience (e.g., Baron (2006)).

The literature review by Rothaermel, Agung & Jiang (2007) summarizes these two categories (see their Figure 7 on p. 761). These theories have only rarely been assessed in a single analysis, and even when examined jointly suffer from a fundamental empirical problem impairing the entire literature on the antecedents of academic entrepreneurship: unmeasured latent commercializability. By this we mean that each scientific discovery has a distinct level of commercial potential, which may be difficult to discern (and is perhaps unclear even to the participants). Indeed, the literature on academic commercialization frequently characterizes academy-originated technologies as "embryonic" (e.g., Jensen & Thursby (2001)), which compounds the difficulty of ascertaining eventual suitability to the commercial market.

Researchers have only rarely attempted to control for latent commercializability. One example is found in Azoulay, Ding & Stuart (2007), who construct such a measure for the life sciences based on keywords assigned by PubMed that overlap with words in patent applications. But even within a set of keywords, there may be vast differences in commercial potential. We instead tackle this confound by building on the Bikard & Marx (2019) method of analyzing "twin" scientific discoveries that arguably have identical commercial potential. We dramatically scale up their effort to include all fields of science over a 60 year period, studying the antecedents of startup commercialization among more than 20,000 twin discoveries. This approach allows us to examine the resource munificence and discovery team composition views on a comparative basis and while taking off the table technology differences (and their latent commercializabilty). Our "twin"-discovery approach therefore mitigate two inference problems plaguing the prior literature. First is the issue of spurious correlations resulting from unaccounted-for differences in commercial potential. Second, even if estimated correlations are not spurious, without controlling for commercial potential it is difficult to discern the degree to which results are due to selection.[1]

Our study illustrates how not accounting for latent commercializability can dramatically alter inferences regarding the antecedents of academic entrepreneurship. We begin with a cross-sectional analysis on a matched sample drawn from the population of published academic findings (more than 42 million academic articles in the Web of Science (WoS), 1955-2017) but *without* controlling for latent commercializability. Through this analysis, we largely replicate prior results confirming the importance of both resource-munificence and discovery-team-composition for entrepreneurial commercialization. When we account for commercial potential by analyzing twin discoveries, however, although we confirm the discovery-team composition view, we find limited evidence for the resource-munificence perspective. Our empirical approach ensures that these differences are not due to different variable definitions or data sources.

Moreover, controlling for latent commercialization refines our understanding of the role of discovery-team composition. Only when examining "twin" discoveries can we conclude that prior results regarding

1. We discuss these issues in the next section. As an example of these inference challenges, consider the finding of Stuart & Ding (2006) that prior association with a professor possessing entrepreneurial experience predicts the focal professor also doing so. This could reflect learning about what it means to be an entrepreneur, as typically interpreted in work on peer effects (Nanda and Sørensen 2010). In the case of academic entrepreneurship, however, it could be due to the influence of a mentor on project selection. That is, researchers intentionally select scientific avenues of inquiry with higher commercial potential. Without being able to account for latent commercializability, it is difficult to tease this mechanism apart.

peer effects in academic entrepreneurship are not driven by selection–for example, mentors pushing their students to choose projects with commercial promise. Namely, *even when considering the same scientific discovery*, scientists with entrepreneurial peers are more likely to commercialize that discovery via a startup. By the same token, although in the cross-section we do not find an association between interdisciplinary research teams and entrepreneurial commercialization, once we account for latent commercializability, we find a robust positive effect. That is, when more interdisciplinary teams of scientists develop the same scientific discovery as less diverse teams, they are more likely to commercialize via startup formation.

We describe latent commercialization-based bias in the technology entrepreneurship literature in Section 2, followed by describing our empirical approach in Section 3. In Section 4 we compare results with and without controlling for latent commercialization. In Section 5, we discuss how these results compare to the existing literature, and a concluding Section 6 reviews our contributions and highlights limitations.

2 Bias in the literature stemming from latent commercializability

In this section, we discuss the classic econometric issue of omitted variable bias (OVB) stemming from unmeasured and/or unobserved latent commercializability of a scientific advance in predicting entrepreneurial commercialization in ordinary least squares (OLS) regression. Consider a "true" regression model of an outcome of interest, $ENTCOMM_i$, which indicates whether the focal scientific advance i was commercialized via a startup. $RESOURCE\_MUNIFICENCE_i$ captures variables proxying commercialization resources potentially available to the authors of the advance, such as the local abundance of potential funding from professional investors such as venture capitalists. $DISCOVERY\_TEAM\_COMPOSITION_i$ measures attributes of the scientific team such as commercial experience. The shadow commercial potential of the advance is captured by $LATENT\_COMMERCIALIZABILITY_i$, and ϵ_i is the error term.

$$\begin{aligned} ENTCOMM_i = \alpha_0 &+ \alpha_1 RESOURCE\_MUNIFICENCE_i \\ &+ \alpha_2 DISCOVERY\_TEAM\_COMPOSITION_i \\ &+ \alpha_3 LATENT\_COMMERCIALIZABILITY_i + \epsilon_i \end{aligned} \tag{1}$$

Now consider that $LATENT\_COMMERCIALIZABILITY_i$ is unmeasured and thus omitted from the specification. Therefore an OLS regression estimates the following equation:

$$\begin{aligned} ENTCOMM_i = \alpha_0 &+ \alpha_1 RESOURCE\_MUNIFICENCE_i \\ &+ \alpha_2 DISCOVERY\_TEAM\_COMPOSITION_i + v_i \end{aligned} \tag{2}$$

Latent commercializability is omitted from the estimated regression and is instead captured within the error term, v_i. Furthermore, there is good reason to believe that latent commercializability is positively related to the outcome variable. To fulfil the conditions of the Gauss-Markov theorem that OLS is the best linear unbiased estimator, the expected value of v_i, conditional on resource munificence and discovery team composition, must equal zero. However, there may be reason to believe that the covariance between latent commercializability and resource munificence (for example), may not be zero.

For example, if a given geographic locale contains many promising researchers and universities with higher potential of producing a commercially-successful product or service, that region may be more

munificent with regard to venture capital funds. Similarly, discovery teams with commercially-experienced members may select projects with more commercial potential, which will induce a positive correlation with the omitted variable, latent commercializability. Because the expected value of v_i, conditional on resource munificence and discovery team composition, is not zero and because the omitted variable is related to the outcome variable, estimates of α_1 and α_2 are biased (the direction of bias as compared to the true α_1 and α_2 depends on the sign on the various covariance relationships).

One remedy is to find a suitable proxy for latent commercializability to include in the regression. Azoulay, Ding Stuart (2007) did so by constructing a measure of "latent patentability" of a faculty member's research (and therefore patenting propensity by academic scientists). They write (p. 600): "While latent patentability typically has been assumed to be unobservable, we are able to devise a patentability score for each scientist in our sample by using keywords in the publications of scientists that have already applied for patent rights as a benchmark for patentable research, and then comparing the research of each scientist in our dataset to this benchmark. Although there is noise in this proxy, it nevertheless quite strongly predicts a patenting event." Controlling for latent commercializability—which is rare in the technology management and entrepreneurship literature—would help address OVB but may introduce other potential biases related to measurement error, precisely because of the noisy measurement.

Our goal is to address latent commercializability-based OVB via a different empirical strategy, one which both directly controls for the technical advance itself and provides natural comparison groups. We examine instances of scientific co-discovery in which there is variation in both the outcome and explanatory variables of the regression model, and by including twin discovery fixed effects (more details in the next section), we aim to sidestep bias stemming from both omitted variables as well as measurement error.[2] To foreshadow our results, we find that addressing omitted latent commercializability in this way changes inference on the importance of various explanatory variables relative to the prior literature.[3]

3 Empirical approach

As noted above, the latent commercializability of scientific discoveries is a critical confound in the literature on academic entrepreneurship. The ideal experiment would involve random matching of researchers and discoveries, which is impractical as few scholars would consent to being assigned projects or colleagues. Instead, we take advantage of the fact that different research teams sometimes make duplicate (or very similar) discoveries. We label these discoveries "twins" as they allow us to hold constant technological differences in shaping startup commercialization. The counterfactual is therefore: if a given scientific advance had been made in two different entrepreneurial ecosystems, is there a tendency for the advance made in the more munificent financial environment to become commercialized by a startup? In this section,

2. This empirical strategy also allows us to improve inference by mitigating the possibility of confounding relationships. Latent commercializability can be a confounding factor in the estimated empirical relationship between discovery team composition and entrepreneurial commercialization. In the estimated (not true) equation, regressing startup formation on observed discovery team characteristics confounds whether team characteristics predict a commercializable discovery or predicts startup commercialization. By holding a discovery constant and relating varied discovery team composition to entrepreneurial commercialization, we can tease these effects apart (we thank an anonymous reviewer for this point).

3. Relative to cross sectional type approaches, other streams of research have also demonstrated that empirical strategies for addressing unobserved variables can overturn conventional results, such as Hsu (2004) in analyzing entrepreneurial affiliation with prominent venture capitalists.

we outline the method for assembling the twins dataset.

We adopt and expand upon a method of identifying twins based on common citation patterns (Bikard and Marx 2019). Citations act as a window into the allocation of credit within the scientific community (Cozzens 1989), so one can infer co-discovery status from papers with distinct authorship but similar citation patterns. Although co-discoveries are uncommon in the social sciences, they are frequent in the "hard" sciences as many research teams are chasing the same scientific frontier. Journal editors may appreciate the opportunity to publish concurrent scientific advances—indeed, twins often appear back-to-back in the same issue of the same journal—as a reaffirmation of the accuracy of the finding. Indeed, Bikard & Marx (2019) verified the method by hiring 10 postdoctoral researchers to manually review dozens of twins that had been identified automatically, with no false positives reported.

We begin by replicating exactly the methodology of Bikard & Marx (2019), finding all pairs of papers that satisfy five conditions: 1) published no more than a year apart; 2) zero overlap between the authors; 3) are cited at least five times; 4) share 50% of forward citations; 5) jointly cited by at least one other paper (i.e. in the same reference list). Our methodology departs from theirs in that instead of limiting our analysis to articles from the top 15 scientific journals between 2000 and 2010, we apply these criteria to the entire Web of Science (WoS) from 1955-2017. Doing so yields a set of potential twin discoveries embodied in 40,392 papers. The next step in the methodology is to determine which of the potential twin discoveries are cited not just jointly (i.e., in the same reference list) but adjacently (i.e., within the same parenthesis). Adjacent citations suggest that forward-citing researchers are unable to attribute the discovery to a single paper, with the references listed within the citation parenthesis receiving co-attribution.

Identifying adjacent citations involves inspecting the text of papers that jointly cite what may be twin discoveries. For the 40,392 potential twin papers, both appear in the reference lists of more than 1.2M papers. Retrieving all such papers is impractical, as many if not most published articles reside behind paywalls and are inaccessible at scale. However, PDFs of many papers are freely available—sometimes in draft form—and have been indexed by Google Scholar (GS). Although GS does not support bulk downloads, over a period of 19 months we were able to retrieve approximately 280,000 publicly-available, non-paywalled PDFs of the 1.2M papers that jointly cited our potential twin discoveries. For 29,257 of the 40,392 potential twin discoveries, we were able to determine whether they were adjacently cited by the PDFs that cited both of them. Of those, we found that 23,851 potential twin papers were indeed cited adjacently.[4] These comprise our population of twin discoveries, which should have similar latent commercial potential.[5] These twin papers reported results from 11,923 twin discoveries. Appendix A provides more detail on the twin discoveries, which hail from more than 3,000 academic institutions in 106 countries and span more than 200 scientific fields.

4. Of the 23,851 twins identified, multiple adjacent citations were found for 62%. This count of adjacent citations is a lower bound, as we could retrieve only 280,000 of the 1.2M papers where both twins are in the reference list. If it had been possible to inspect all 1.2M papers, we likely would have found multiple adjacent citations for more twins. In Table 5, we drop the twins established via a single adjacent citation, yielding similar results.

5. Twin discoveries are not randomly distributed, however, and so in our cross-sectional empirical comparisons (which do not control for latent commercializability), we undertake a matching strategy to bring the twin and non-twin samples into better balance along key observable characteristics.

3.1 Dependent variable: entrepreneurial commercialization of scientific discoveries

Our dependent variable indicates whether academic researchers commercialize their discoveries via a startup. To our knowledge, a large sample of academic scientific discoveries commercialized via startups has not been previously assembled. Several studies of technology transfer have tracked out-licensing or other forms of commercialization more generally, not necessarily via new venture formation (see Rothaermel, Agung & Jiang (2007) for a review). There have also been academic institution-specific studies of new venture formation (e.g., Kenney & Goe (2004); O'Shea et al. (2005)) as well as sector-specific studies of commercial science, most notably in the biotechnology industry (e.g., Zucker, Darby & Brewer (1998); Stuart & Ding (2006)). Our aim, however, is to identify entrepreneurial commercialization of scientific discoveries at scale, spanning academic institutions, industrial sectors, and geography. Aside from the benefit of algorithmically assembling a large empirical sample, our method allows us to directly trace new ventures all the way back to a particular scientific advance, a feature also novel to the literature.

We measure entrepreneurial commercialization in two ways. First, via patent-paper pairs ("PPPs") (Murray 2002) where the patent is assigned to an entrepreneurial venture. The premise is that while scientific publications are the currency of academia, patents and their associated legal protection are valued much more in the commercial domain. Our effort is aimed at identifying patents granted to entrepreneurial ventures that cover the same or similar scientific advance in which there is overlap between inventors and authors. We start by finding the subset of academic discoveries that are cited by patents (Marx and Fuegi 2020) and check for overlap between the authors of the paper and the inventors named on the patent. Article authors and patent inventors are compared individually, with an overall match score computed according to a) whether the surname is an exact versus fuzzy match; b) frequency of the surname in the WoS and the patent corpus; and c) whether the middle initial matches (more details are provided in Appendix B). A weighted average of author/inventor overlap is computed to yield an overall article/patent match score.[6] However, not every patent-paper pair represents entrepreneurial commercialization. For example, one or more scientists on a paper may cooperate with an established firm to commercialize the discovery. We thus subset the list of PPPs to those assigned to startups, as determined from VentureSource and Crunchbase.

Our second method involves U.S. Small Business Innovation Research (SBIR) grants. The SBIR program is targeted at encouraging "domestic small businesses to engage in federal research and research development that has the potential for commercialization" and has awarded non-dilutive funding in excess of $45B since the program was initiated in 1982 (www.sbir.gov/about). We interpret pursuing SBIR funds as an indicator of commercialization aspirations. Note that the SBIR channel of identifying commercialization attempts does not rely on observed patenting: this may be an important complement to the PPP measure, as Fini, Lacetera & Shane (2010) suggest that only about a third of businesses started by academics are based on patented inventions. Moreover, the literature's reliance on patent data is likely related to the fact that our understanding of technology commercialization heavily relies on the biotechnology industry (see, for example, the discussion in Hsu (2008)), as patents are well-understood to be

6. If the authors of an article have an identical overlap score with the inventors on multiple patents, ties are broken in two steps. First, the PPP closest in time is retained. Second, if two patents in the same year form pairs with the same paper, we further resolve ambiguity via cosine similarity between the abstract of the article and the summary text of the patent.

important as an appropriation method in that industry (e.g., Levin, et al. (1987)). As with patent-paper pairs, we calculate the pairwise overlap between scientists on a focal article and either the primary contact or principal investigator of SBIR awards up until five years after the publication of the article. If multiple SBIR awards have identical author-overlap scores, we break ties with temporal proximity.

Overall, we find 139 academic articles that were commercialized via PPPs assigned to startups and 89 that were commercialized via SBIR awards, for a total of 228 entrepreneurial commercialization events. Appendix B also provides validation of the measure, confirming via web research a stratified random sample that both PPPs and overlapping SBIR grants truly reflect instances of a startup commercializing an academic discovery with the involvement of one of the original scientists. In short, we verified 20 out of 20 of the PPP-based commercialization events, and 19 out of 20 SBIR-based events.

3.2 Explanatory variables

Our explanatory variables fall into the aforementioned categories of resource munificence and discovery-team composition. Resource munificence is often tied to geography (Samila and Sorenson 2011), so we constructed a lagged count of venture capital investments in the same postal code as the focal article. Resources also often accrue to high-status actors, so our second and third measures of munificence reflect the prestige of the discovery team and their associated institutions (Stuart, Hoang, and Hybels 1999). Each of these variables is calculated as a count of publications (per author, or per institution) in the same scientific field as the focal paper. WoS assigns each article to one of 251 scientific fields.[7]

Regarding discovery team composition, a first variable measures the interdisciplinarity of the scientists. This is calculated as one minus the Herfindahl-Hirsch index of scientific fields for articles written by the authors. If all scientists on the focal article published all of their papers in the same scientific field, this variable is zero. A second explanatory variable measures whether the previous collaborators of authors on the paper include a 'star commercializer.' This variable is reminiscent of Stuart & Ding's (2006) measure of the number of prior collaborators who served as founders or advisory-board members of startups that filed for an IPO, but our measure differs in three ways. First, we measure involvement with early-stage ventures, not just those that complete an IPO. Second, instead of summing all instances of entrepreneurial involvement, we focus on "star" serial entrepreneurs (above the 75th percentile of entrepreneurially-commercializing academic scientists in the year of the scientist's most recent collaboration (similar results are obtained at the 50th or 90th percentile in Table 4)). Third, we check whether any scientist on the discovery team had previously collaborated with a star. Additional characteristics of star commercializers are available in Appendix C. As a third team-composition covariate, we control for whether any of the authors on the paper is herself a star commercializer.

7. For institutions in North America, we also have technology-transfer related variables from the Association of University Technology Managers and compute models limited to institutions where such variables are available. However, because the AUTM data rely on respondent survey responses which are self-reported and because of the limited (domestic) coverage of the data only among some association members, we do not report these models.

3.3 Empirical specification

Following epidemiological twin studies (Carlin et al. 2005), we estimate the likelihood of commercialization using fixed effects for articles reporting a twin discovery. The regression equation is:

$$\begin{aligned} ENTCOMM_{ij} = \alpha_0 + \alpha_1 RESOURCE\_MUNIFICENCE_{ij} \\ + \alpha_2 DISCOVERY\_TEAM\_COMPOSITION_{ij} + \alpha_3 X_i + \gamma_j + \epsilon_{ij} \end{aligned} \tag{3}$$

In the specification, j represents the twin discovery and i represents an article reporting the twin discovery. $ENTCOMM_{ij}$ captures whether the focal article was commercialized by a startup. Variables related to local venture capital investments and the prestige of the institution and discovery team are captured in $RESOURCE\_MUNIFICENCE_{ij}$. $DISCOVERY\_TEAM\_COMPOSITION_{ij}$ variables measure the interdisciplinarity of the scientific team and whether the scientists on a given article had previously collaborated with a "star" entrepreneurial commercializer. X_i represents controls for the number of authors and count of citations from industry patents. γ_j is a fixed effect for the twin discovery, and ϵ_{ij} is the error term. We estimate this equation using linear probability models with robust standard errors. Following Beck (2018), in the robustness checks of Table 5 we also estimate conditional logit models, which exclude any twin discovery where neither is commercialized, with similar results.

4 Results

We present results in two steps. Before proceeding to analyze the entrepreneurial commercialization of academic science via our twins methodology, we present cross-sectional estimates drawing from the 42M+ scientific publications contained in the Web of Science. Several findings from the past literature can be replicated in this setup, which does not account for latent commercializability. Then we contrast the cross-sectional results with those obtained via the twin-paper setup described above. Our motivation for doing this is twofold. First, as noted above despite hundreds of papers on related topics, resource-munificence and discovery-team composition factors have only rarely been jointly considered. Second and more importantly, to the extent that our findings using the twins methodology differ from prior work, one might suspect that this is due to variable construction or the scope of the data analyzed. As noted above, past work has frequently analyzed patented inventions, invention disclosures, or projects from a small number of universities. We are largely able to replicate prior results using a sample drawn from the universe of scientific publications. However, when addressing technology and latent commercializability differences via the twins method, the magnitude and sometimes the sign of results change materially.

4.1 Cross-sectional results

The Web of Science contains more than 42 million academic articles. Among these are 11,340 instances of entrepreneurial commercialization. Of those, 8,361 were found via patent-paper pairs and 2,966 via SBIR grants.[8] One approach would be to conduct a cross-sectional analysis of the *entire* WoS, but differences

8. 17 were in both categories. This modest overlap probably arises from the specific set of inclusion criteria for the PPP and SBIR measures, as described in Section 3.1. To verify that our measures are capturing what they should, we examine if there are false negatives among the SBIR sample of firms with a patent in which our algorithm should have flagged as

between such results and those obtained using the twins methodologies might be ascribed to differences in the samples. Indeed, descriptive statistics segmented by twin versus non-twin observations (Panel A of Table 1) reveal substantial differences between the twins subsample and the larger population. The distribution of our measures of both resource munificence (*author prestige*, *institution prestige*, and natural-log transformed *VC investments in the postal code*) and discovery team composition (*interdisciplinarity* and *prior work w/star commercializer*) significantly differ between the two sub-samples. In particular, the average value for the twins sub-sample is higher than that of the non-twins for four out of these five variables (all except *interdisciplinarity*).

Table 1 about here

This imbalance suggests that the data generation process for twins is different than that of the rest of the WoS; perhaps twin discoveries are more important than the average scientific publication. Therefore, we pre-process the WoS data using Coarsened Exact Matching in order to balance covariates with the twins sample: *author prestige*, *institution prestige*, natural-log transformed *VC investments in the postal code*, *interdisciplinarity*, and whether the authors had prior affiliation with a 'star' commercializer (*prior work w/star commercializer*). Continuous variables were segmented into ten bins each for purposes of matching. As a result, each observation was categorized into one of 559 "strata" (i.e., combinations of bins for purposes of matching with the relevant variable from the twin sub-sample). For strata where both twin and non-twin observations were categorized, one non-twin observation was selected (randomly) for each twin in the strata. Of the 23,851 twins, 19,245 were matched to a non-twin.[9]

The resulting sample balance is shown in Panel B of Table 1. All of the variables are now much better balanced in value. In essence, the matching procedure trades off a (much) larger sample size to match the twin sample on key observables.

We use the resulting matched sample to conduct a cross-sectional regression analysis in Table 2. Note that these regressions do *not* control for latent commercializability and therefore are susceptible to the omitted variable bias previously discussed. All specifications include controls for number of paper authors, as well as counts of citations from industry patents, as well as fixed effects for year, country, and scientific field. The estimating equation for Table 2 is therefore:

$$\begin{aligned} ENTCOMM_{ij} = \alpha_0 + \alpha_1 RESOURCE\_MUNIFICENCE_{ij} \\ + \alpha_2 DISCOVERY\_TEAM\_COMPOSITION_{ij} + \alpha_3 X_i + \eta_i + \zeta_i + \theta_i + \epsilon_{ij} \end{aligned} \quad (4)$$

where η_i is a fixed effect for year of the article, ζ_i is a fixed effect for the country of the article's corresponding author, and θ_i is a fixed effect for the scientific field of the article as assigned by WoS.

Columns (1-6) of Table 2 follow the literature by examining each explanatory variable independently. We find positive, statistically significant effects at the $p<0.01$ level for two measures of resource munificence: (*author prestige* and *VC investments in the postal code*), as well as for two measures of discovery

a PPP. We start by (exact) matching our sample of SBIR winners with patent assignee names. About half of the SBIR awardees did not have patents. On the subsample which did have a patent, we undertake a series of manual checks to see if we missed any overlaps with the PPP route. We removed any observations with inconsistent geography, time, or individual name information. On the remaining observations, we hand-checked a random 10% sample for misclassifications by inspecting the non-patent references using Google Patents. Through this process, we did not find any misclassifications.

9. In some cases, more than one twin paper were matched to the same non-twin paper, so the number of non-twins in Panel B of Table 1 is lower than the number of twins.

team composition (*prior work w/star commercializer* and *star commercializer on paper*). These results are consistent with that of prior work. When all covariates are considered jointly in column (7), we find similar effects as the individual regression, with the exception of the *author prestige* variable, which is less precisely estimated.[10]

Table 2 about here

The results suggest, consistent with prior research, that local financial capital facilitates commercialization, supporting the resource munificence view. Moving to discovery-team composition factors, we likewise confirm Stuart & Ding's (2006) findings regarding the strong role of both having a star commercializer either on the discovery team or among one of the team's prior collaborators. By replicating prior results from the literature based on a matched sample drawn from the full population of academic articles since the 1950s, we hope to alleviate concerns that the differences we find in the following section via our twin-discovery method are due to either data or variable construction differences. Next, we show that the results differ when controlling for latent commercialization.

4.2 Twin-discovery results

The above analysis embeds latent commercializability effects in the error term, yielding OVB due to the theoretical correlation between latent commercializability and entrepreneurial commercialization as well as the correlation between latent commercializability and both groups of independent variables of interest. Our approach to overcoming OVB is to constrain our sample to just the scientific twins, including twin fixed effects in each specification. We begin in Table 3 by emulating the specification structure from Table 2, limiting the sample to articles that report twin discoveries, and replacing fixed effects for year, country, and scientific field with fixed effects for the twin discovery as in Equation (3).

Table 3 about here

In columns (1-3), we reassess the role of resource munificence using the twins analytic strategy. As compared to column (1) of the cross sectional regression in Table 2, when accounting for latent commercializability in column (1) of Table 3, *author prestige* does not materially impact entrepreneurial commercialization. Similar to the cross-sectional result, *institution prestige* remains insignificant. In addition, we find that the entrepreneurial commercialization of academic science is not explained by the number of local venture-capital investments when accounting for latent commercializability (compare columns 3 of Tables 3 vs 2). Although we discuss in more detail how our results relate to findings from the literature in Section 5, one explanation for the difference in the local VC effect may be selection of researchers with higher commercial potential projects into particular geographic regions.

In columns (4-6) of Table 3, we apply the twin-discovery methodology to assess how controlling for latent commercializability impacts the discovery team composition effects (relative to the cross-sectional estimates). When interpreting these measures, we remind the reader that our variables do not necessarily reflect the composition of the founders of the startup but rather the team of scientists that pioneered

10. In unreported results, we estimate the above equation on the entire Web of Science, with similar results.

the original scientific finding. Under our definition, one or more of the authors was indeed involved with the startup. Column (4) indicates that discoveries where the scientific team is more interdisciplinary are somewhat more likely ($p<0.10$) to be commercialized by startups than when the discovery team is homogeneous with respect to scientific field. Column (5) shows that discoveries are more likely to be commercialized via startups when one of the authors is a star commercializer. In column (6), we find similar confirmation for discovery teams where one of the authors previously worked with a star commercializer. All resource-munificence and discovery-team-composition covariates are included in column (7), which strengthens statistical significance on the estimated coefficient for interdisciplinarity to the $p<0.05$ level.[11]

Using estimated coefficients from column (7), a one-standard-deviation increase in interdisciplinarity (0.27) corresponds to a 2.7% increase in the likelihood of commercialization by a startup. The presence of a star commercializer among the scientists' past collaborators is associated with a 4.1 percentage-point increase in commercialization, and having a star commercializer among the authors themselves predicts a 9.5 percentage-point rise in the likelihood of commercialization. Thus we obtain quite different results when controlling for latent commercializability via our twin-paper strategy.

4.3 A deeper dive into discovery team composition effects

A key difference between the cross-sectional and twin regressions is the emergence of a role for interdisciplinarity in the commercialization of academic science. In Table 4, we dig deeper into the nature of interdisciplinarity and "stars." Interdisciplinarity may take a number of forms. An interdisciplinary team could be composed of several specialists from different scientific fields, possibly with one boundary spanner. Alternatively, an interdisciplinary team might be composed of generalists who are themselves interdisciplinary. In columns (1-5) of Table 4, we explore which forms of interdisciplinarity matter for entrepreneurial commercialization. In column (2), we replace the *interdisciplinarity* variable from Table 3 (shown in column (1) for comparison) with an indicator for above-mean interdisciplinarity. The coefficient remains positive, though weaker. In column (3), we count the number of unique primary scientific fields among the authors on the paper. By "primary" we mean the scientific discipline in which each author publishes most often. The positive, statistically-significant estimate of the associated coefficient suggests that having scientists from a variety of scientific fields is important, not just having a set of scientists from the same discipline who also work relatively often in other areas. The magnitude of the estimated coefficient on the simple count of scientific fields represented is smaller than the more subtle measure of interdisciplinarity among all authors, however. This suggests that merely having more scientific fields represented does not fully explain the findings in column (1).

Table 4 about here

Although column (3) suggests that specialization is useful, column (4) clarifies that the optimal discovery team for commercialization requires more than a team of specialists. The covariate in this column counts the number of scientists who fully specialize (i.e., all of their work is published in a single scientific field). If specialization alone were critical to the commercialization process, we might expect this

11. In unreported results, weremove fixed effects for the twin discovery and recover results reminiscent of Table 2. In particular, the coefficient on venture capital proximity is positive and precisely estimated.

coefficient to be significant, but it does not appear so. Alternatively, perhaps it is the case that an ideal configuration would combine a set of specialists with a boundary spanner. In column (5), we calculate each scientist's individual level of interdisciplinarity and then enter as a covariate the difference between the most interdisciplinary scientist and the mean of the team. The negative coefficient suggests that having one scientist with much more interdisciplinarity than most does not facilitate commercialization. Taken together, columns (2-5) suggest that a well-rounded team of well-rounded scientists may be particularly effective in spurring entrepreneurial commercialization. As we discuss further in section 5.3, this result is consistent with Baron's (2006) claim that "opportunity recognition can be enhanced by providing potential entrepreneurs with a very broad range of experience... the broader this experience... the more likely the entrepreneurs will be to perceive connections between seemingly unrelated events or trends."

We similarly explore whether alternative definitions of *star commercializer* and *prior work w/star commercializer* overturn the positive estimated effects from Table 3. In columns (6-7) of Table 4, we see that replacing those measures with indicators for above median values (column 6) or being in the top 10% of each distribution (column 7) also results in positive and statistically-significant estimated coefficients.

The remaining column (8) of Table 4 verifies that it is the presence of a star *commercializer* among the scientists' past collaborators that explains the patterns in Table 3 and not simply an association with a prominent researcher. We replace the star commercializer variable with an indicator for a star scientist—those whose citation count per article (in a five-year window following publication) was in the 99th percentile—among one's past collaborators. Following Zucker et al. (1998), we would expect this coefficient to be statistically significant, but it is imprecisely estimated here. We conclude that star commercializers, as opposed to star scientists, facilitate the entrepreneurial commercialization of science.

4.4 Robustness and placebo tests

Finally, we establish the robustness of the findings thus far to a variety of specifications. Column (1) of Table 5 repeats column (7) of Table 3 for comparison. Column (2) re-estimates column (1) using conditional logit (see Beck (2018)). Because the maximum likelihood estimator drops groups without variation in the dependent variable, the fixed effects for each twin discovery reduces the number of observations. Statistical significance is reduced somewhat for the interdisciplinary result ($p<0.08$) but otherwise resembles column (7) of Table 3. In column (3), we present the OLS specification analog to the logistic regression from column (2) by manually limiting observations to the set of twin discoveries with variation in the outcome variable. When doing so, OLS results closely resemble logit.

As noted above, we were able to check for adjacent citations (future citations which reference our candidate twin papers within the same parentheses) in only 280,000 of the 1.2 million papers that jointly cited potential twins. Although for 38% of our twins we found only a single adjacent citation, that figure may be understated as we could not inspect three-fourths of the co-citing PDFs. We thus check that our results are robust among twins that we *can* confirm were adjacently cited multiple times. Column (4) confirms the results are similar, with a slightly smaller t-statistic for the interdisciplinarity coefficient.

Finally, in column (5) of Table 5, we perform a placebo test by randomly generating values for the dependent variable. Doing so yields no statistical significance on any covariates. In unreported results, this placebo test also fails if the distribution of the randomly-generated dependent variable matches that

of the actual dependent variable (i.e., much less than 1% of papers are commercialized by startups).

Table 5 about here

Next, we check that the null results regarding the effects of resource-munificence factors are not driven by our particular variable construction. To this end, we generate "specification maps" (King, Goldfarb, and Simcoe 2019). In Figure 1 we graph the estimated *author prestige* coefficient using the specification from column (7) of Table 3, but substituting various alternative ways of measuring the construct: logged and unlogged count of papers, a count of citations instead of papers, average journal impact factor (JIF) for authors on the paper, and average eigenfactor (a measure of the importance of journals that cite the author's papers). All estimated coefficients are not different than zero (statistically). Figure 2 follows a similar logic in exploring whether the estimated zero effect of *institutional prestige* depends on its exact measurement, finding no statistically significant effects using any alternative measure.

In Figure 3, we investigate whether the non-effect of local venture capital availability is an artifact of the measure of the construct. We plot the estimated coefficient (and 95% confidence interval) based on the specification of column (7) of Table 3, but measuring VC investment in alternative ways: logged investments in the postal code; unlogged investments; replacing the count of investments per postal code with a count of liquidity events (i.e., IPOs or acquisitions); replacing the data source of investment counts using VenturExpert instead of Crunchbase; and measuring local VC investment at the Core-Based Statistical Area (CBSA) level instead of the postal code. All of these measures also result in an estimated VC effect not statistically different from zero.

In Figure 4, we create another specification map (King, Goldfarb, and Simcoe 2019) that plots the regression coefficient of *Ln same-postalcode investments (CB)* using various specifications. The first row shows the estimated coefficient and 95% confidence interval from column (7) of Table 2, which is positive and statistically distinct from zero. In the second row, we replace the matched subset of the 42+M papers in WoS with the twin papers (i.e., the same sample as in Table 3), but *without* twin fixed effects. The estimated coefficient is still statistically distinct from zero, indicating that the null result in Table 3 is not simply due to the sampling population. The third row resembles the second row, except that the sample is reduced (at random) to only one paper from each twin discovery such that it is not possible to use fixed effects for the twin discovery. In this specification, with slightly less than half the data points, the coefficient is still positive and significant. Only in the fourth row, where we include twin fixed effects (column (7) of Table 3) and therefore account for latent commercializability, does the estimated coefficent on *Ln same-postalcode investments (CB)* become indistinct from zero.

In Appendix D we examine whether the results from Table 3 are heterogeneous. In summary, we find that the role of stars and interdisciplinarity depends on geographic location, discipline, and time period.

5 Calibration with the literature and future directions

We now discuss how the population-level, cross-sectional estimates contrast with those from the twin study, and how these results compare with those reported in the literature.

5.1 Entrepreneurial ecosystem resource munificence

The munificence of resources required to commercialize entrepreneurial discoveries has been a frequent subject of inquiry. For example, Zucker, Darby, & Brewer (1998) report that U.S. states with more academic scientists who have outsized academic output ("stars") are home to more biotechnology startups, particularly in the nascent phase of industry development. Their study does not establish direct linkages between startups and academic scientists, but their findings have generally been interpreted to suggest that the localized knowledge of highly productive scientists may be important in the geography of entrepreneurial commercialization. Such an interpretation is consistent with what we see in the matched cross-sectional results in column 1 of Table 2. However, when accounting for latent commercialization via our twins methodology (in Table 3 and in the robustness results of Figure 1), we no longer find support for the role of prominent scientists.

Similarly, an extensive literature has found a connection between the local munificence of venture capital and the founding of new firms (Samila and Sorenson 2011). We see a similar association in the cross-sectional analysis of Table 2 but not when we employ our empirical approach of controlling for latent commercializability in Table 3. The null results are confirmed in a variety of robustness tests using alternative measures for the local VC investment construct in Figure 3, as well as alternative empirical specifications in Figure 4. The fact that we fail to find any connection once applying the twins methodology raises the question of whether this effect extends to *academic* entrepreneurship. It could instead be that discoveries in close proximity to sources of capital simply have more commercial potential, raising the possibility that commercially-minded academic scientists select into regions with locally-available financial capital.

Note that both the Zucker, Darby & Brewer (1998) and Samila & Sorenson (2011) papers are set in the US and (largely) in the time period before the year 2000. While it is difficult to directly compare our twins results on commercializing science with these papers (the outcome variables and the samples differ), we do not find evidence for the local knowledge capital or VC effect in our analysis of sample heterogeneity (Appendix D). This suggests that further research on the role of resource munificence while taking into more careful account the nature of technical opportunities produced in the local geography, especially as related to academic entrepreneurship, may be warranted.

5.2 Discovery team composition: "star" commercializers

Whereas we failed to find support for resource munificence when accounting for latent commercialization, we affirm the role of star commercializers (Stuart and Ding 2006) both in the matched cross-sectional analysis and when using the twins methodology. The positive effects are robust to a variety of alternative ways of measuring star commercializers (as seen in Table 4). In addition to reconfirming prior findings, by controlling for latent commercialization we show that peer effects are not limited to project selection. That is, one might suspect that prior association with a star commercializer might lead a focal researcher to pick research projects with greater commercial potential. This may well be true, but our results show that *even given the same scientific discovery*, those with exposure to entrepreneurial peers are more likely to commercialize their research.

Our findings are consistent with broader work on peer effects, such as Nanda & Sorenson (2010), who use Danish data from 1980-1997 and find a positive immediate workplace peer effect on entrepreneurial entry. While it is difficult to directly compare our results to these studies due to different samples (among other things), the operative mechanisms behind the positive peer effects are demonstration effects ("if I can be an entrepreneur, you can as well") and information/resource effects ("in order to be credible to investors in this space, you have to have more than a product prototype").

It is important to stress that our results should not be interpreted as causal evidence for peer effects in academic entrepreneurship. Although it is important to hold constant the nature of the discovery, we do not address the endogeneity of discovery team formation. Perhaps the original scientist recruits a star commercializer to join the discovery team if s/he senses a startup opportunity. We know who the corresponding author is on each paper but not the order in which the team was *assembled* (not just order of authorship) in order to rule out this possible alternative. What we can rule out by controlling for latent commercializability is that project selection was influenced by prior exposure—i.e., pushing scientists toward projects that are more applied and thus have greater commercial promise.

Moreover, Lerner & Malmendier (2013), using randomized graduate business student section assignments between 1997-2004, find peer effects dampen entrepreneurial entry (but that those who enter are more successful, suggesting that peers may provide information allowing focal individuals to "properly" assess their prospects). Their negative peer effect seems to stem from a "screening" function which may serve as a check to would-be entrepreneurs who may not have realized the many obstacles to successful entrepreneurial entry. While we do not study commercialization performance, our heterogeneity results (Appendix D) in comparison to this study suggests that perhaps industry conditions outside of the biotechnology and life sciences may curtail the role of (star commercializer) peer effects. Consistent with the Lerner & Malmendier (2013) study, we do not find a general peer effect outside of the biotechnology and life science contexts. There may be less of a compelling peer screening function outside of the specialized expertise in the life and health sciences.

Although the effects of peers on (academic) entrepreneurial starts is therefore not settled, especially as there are more than a few differences in the research contexts across the studies, we believe that a fruitful path forward is tying peer effects to the specific entrepreneurial opportunity. Note that in contrast to the prior peer effects literature which treats such opportunities as unspecified and unmeasured, our study begins the process of specifying the (scientific) advance giving rise to a potential entrepreneurial opportunity. At the same time, it will be important to pay particular attention to possible differences in the peer effect process in the life and health sciences as compared to other scientific sectors. Such an empirical strategy holds the promise of revealing more on how and why the nature of peer effect interaction matters for entrepreneurial commercialization.[12]

12. One difference between the setting of prior studies versus academic scientists is that faculty undertaking entrepreneurial ventures often remain in their positions. For example, Professor Robert Langer, whose research has spawned over 40 startups, never left MIT; instead, his associates took the lead in commercialization efforts.

5.3 Discovery team composition: interdisciplinarity

Finally, we consider the interdisciplinary nature of discovery teams. Interdisciplinarity has been frequently studied (Leahey, Beckman, and Stanko 2017), though rarely in the context of entrepreneurship. The most relevant articles of which we are aware are Berkovitz & Feldman (2011) and Kotha, George, & Srikanth (2013), both of which examine the licensing of university invention disclosures as opposed to startup formation. These authors find that more departments spanned by the discovery team (which they interpret as coordination costs) is negatively associated with a lower likelihood of licensing (though the effect reverses for a squared term indicating a curvilinear effect). Our matched cross-sectional results do not yield a significant coefficient on the *interdisciplinarity* variable.

When applying the twins methodology, however, the sign of the estimated coefficient on interdisciplinarity becomes positive (Table 3, column 7). That is, when a more interdisciplinary team develops a highly similar invention as a less interdisciplinary team, it is more likely to commercialize via startup formation. By controlling for latent commercialization, we take off the table the role of coordination costs and project selection. Having overcome these costs, there are at least two reasons to think that interdisciplinary teams are more likely to commercialize a given discovery. First, a more diverse set of perspectives among the original scientists may improve opportunity recognition due to varied inputs. Baron (2006, p.17) claims that "opportunity recognition can be enhanced by providing potential entrepreneurs with a very broad range of experience... the broader this experience... the more likely the entrepreneurs will be to perceive connections between seemingly unrelated events or trends." Similarly, Shane & Venkataraman (2000) argue that heterogeneity may "give rise to different entrepreneurial conjectures." Second, a more diverse set of scientists may have broader networks than those who all work in the same field (Hills, Lumpkin, and Singh 1997). Networks may amplify the informational advantage mentioned above, or they may lead to sources of human or financial capital.

Our results regarding discovery team composition while holding constant latent commercializability thus help to reconcile the coordination costs of interdisciplinary work with the potential for enhanced opportunity recognition and resource assembly by a cross-disciplinary team. Our results on heterogeneity (Appendix D) suggest that the predictive role of interdisciplinary discovery teams of academic entrepreneurship is strongest in the U.S., in the non-biotechnology life science sector, and in the pre-2000 time period. The deeper dive into the forms of interdisciplinarity that matter most likewise suggest that the coordination costs of disciplinary-focused researchers are offset by discovery-team members who are themselves interdisciplinary.

As in our prior discussion on discovery team composition, however, the interdisciplinarity results should not be interpreted as causal. In general, future research may delve more deeply into the process of scientific team formation. Boudreau et al. (2017) suggest that there are search frictions associated with the process of finding scientific collaborators. In a field experiment context, they found that randomization in research funding information session colocation among researchers had a substantial (75%) boost in the likelihood that author dyads would submit collaborative proposals. Results like these suggest that a deeper understanding of the antecedents of discovery team formation will be helpful in better understanding how discovery team composition more generally impacts academic entrepreneurship.

6 Discussion & Conclusion

Given interest in translating academic science into commercial products, including via startups, improving our insight into the antecedents of this process is essential. Prior work has yet to account for differences in the inherent commercial potential of scientific discoveries, however, which may result in spurious inferences through omitted variable bias. Based on an algorithmic approach, we assemble a large sample (about 20,000) of scientific co-discoveries, which allows us to hold constant the scientific advance (and therefore addresses the confound of latent commercializability). As compared to a matched cross-sectional empirical strategy which does not address the issue of latent commercializability, we demonstrate that the magnitude and sometimes even the direction of estimated effect can differ. The scientific-twins approach is aimed at holding constant "nature" to focus on the effect of "nurture" or environmental effects, as is the case with human twin studies.

Focusing on two broad classes of mechanisms—resource munificence and discovery-team composition—we confirm the importance of both in a cross-sectional analysis of a matched cross-sectional analysis drawn from all academic articles in the Web of Science from 1955-2017 while *not* accounting for latent commercial differences. However, when controlling for these differences via our twin-discovery approach, we no longer find empirical support for the resource munificence view, most notably the effect of local venture capital investment activity. It may be that selection by researchers with a commercial disposition into prominent institutions or resource-rich geographies is a better explanation for existing findings. Regarding discovery-team composition, however, we both reaffirm and refine existing findings. In both matched cross-sectional and twin-discovery analyses, we find a strong connection between exposure to peers with entrepreneurial experience. In controlling for latent commercialization, unlike prior literature we can state that these peer effects are not driven solely by project selection. Our results also revisit prior findings regarding interdisciplinarity: although in the cross-section we find no effect of interdisciplinary, when controlling for latent commercializability, we observe that interdisciplinary teams are in fact more likely to commercialize.

A limitation of our methodology is that we may not capture scientific commercialization by a startup that licenses or otherwise appropriates the discovery *without* involvement from the original scientists. In addition, as previously noted, our results should also not be interpreted as causal, as team composition is of course not randomly determined. One selection effect could be the unobservability of author teams that did not successfully publish their paper. This would impact the possible censoring of observed "twin" discoveries, especially if the main reason why a given paper is not published is because journal editors decide that the focal paper is not novel given an existing paper already published or accepted for publication in the literature. If author teams of these censored papers are equally distributed by interdisciplinarity and association with star commercializers, this would not present a problem. If, on the other hand, such unobserved paper author teams are much more likely to be uniform with regard to disciplinary background and less likely to have a star commercializer on the author team, then our results may be biased upwards. While we do not think this is likely, the issue illustrates a broader interpretational point associated with our methodology: we take the process generating observed scientific twins as given (and therefore exogenous to our study). As we suggest in the prior section, our findings suggest a fruitful avenue

for future research would be to better understand the antecedents of both discovery team composition as well as the scientific co-discovery. In addition, we believe that there is ample opportunity to examine not just outcomes related to the act of entrepreneurial commercialization itself, but also any number of subsequent venture outcomes and milestones. Doing so would extend our understanding of the lifecycle from scientific advance to commercial impact.

References

Azoulay, Pierre, Waverly Ding, and Toby Stuart. 2007. "The determinants of faculty patenting behavior: Demographics or opportunities?" *Journal of Economic Behavior & Organization* 63 (4): 599–623.

Baron, Robert A. 2006. "Opportunity recognition as pattern recognition: How entrepreneurs "connect the dots" to identify new business opportunities." *Academy of Management Perspectives* 20 (1): 104–119.

Beck, Nathaniel. 2018. "Estimating grouped data models with a binary dependent variable and fixed effects: What are the issues." *arXiv preprint arXiv:1809.06505.*

Bercovitz, Janet, and Maryann Feldman. 2011. "The mechanisms of collaboration in inventive teams: Composition, social networks, and geography." *Research Policy* 40 (1): 81–93.

Bikard, Michaël, and Matt Marx. 2019. "Bridging academia and industry: How geographic hubs connect university science and corporate technology." *Management Science.*

Boudreau, Kevin J, Tom Brady, Ina Ganguli, Patrick Gaule, Eva Guinan, Anthony Hollenberg, and Karim R Lakhani. 2017. "A field experiment on search costs and the formation of scientific collaborations." *Review of Economics and Statistics* 99 (4): 565–576.

Carlin, John B, Lyle C Gurrin, Jonathan AC Sterne, Ruth Morley, and Terry Dwyer. 2005. "Regression models for twin studies: a critical review." *International Journal of Epidemiology* 34 (5): 1089–1099.

Cozzens, S. 1989. "What do citations count? The rhetoric-first model." *Scientometrics* 15 (5-6): 437.

Cyranoski, D., N. Gilbert, H. Ledford, A. Nayar, and M. Yahia. 2011. "The PhD factory: The world is producing more PhDs than ever before. Is it time to stop." *Nature* 472 (7343): 276.

Fini, Riccardo, Nicola Lacetera, and Scott Shane. 2010. "Inside or outside the IP system? Business creation in academia." *Research Policy* 39 (8): 1060–1069.

Haltiwanger, John, Ron S Jarmin, and Javier Miranda. 2013. "Who creates jobs? Small versus large versus young." *Review of Economics and Statistics* 95 (2): 347–361.

Hayter, Christopher S, Roman Lubynsky, and Spiro Maroulis. 2017. "Who is the academic entrepreneur? The role of graduate students in the development of university spinoffs." *The Journal of Technology Transfer* 42 (6): 1237–1254.

Hills, Gerald E, G Thomas Lumpkin, and Robert P Singh. 1997. "Opportunity recognition: Perceptions and behaviors of entrepreneurs." *Frontiers of Entrepreneurship Research* 17 (4): 168–182.

Hsu, David H. 2004. "What do entrepreneurs pay for venture capital affiliation?" *The Journal of Finance* 59 (4): 1805–1844.

———. 2008. *Technology-based entrepreneurship, in: The handbook of technology and innovation management, S. Shane, ed.* John Wiley & Sons.

Hsu, David H, and Tim Bernstein. 1997. "Managing the university technology licensing process: Findings from case studies." *Journal of the Association of University Technology Managers* 9 (9): 1–33.

Jensen, Richard, and Marie Thursby. 2001. "Proofs and prototypes for sale: The licensing of university inventions." *American Economic Review* 91 (1): 240–259.

Kenney, Martin, and W Richard Goe. 2004. "The role of social embeddedness in professorial entrepreneurship: a comparison of electrical engineering and computer science at UC Berkeley and Stanford." *Research Policy* 33 (5): 691–707.

King, Andrew A, Brent Goldfarb, and Timothy Simcoe. 2019. "Learning from testimony on quantitative research in management." *Academy of Management Review,* no. ja.

Kotha, R., G. George, and K. Srikanth. 2013. "Bridging the mutual knowledge gap: Coordination and the commercialization of university science." *Academy of Management Journal* 56 (2): 498.

Leahey, Erin, Christine M Beckman, and Taryn L Stanko. 2017. "Prominent but less productive: The impact of interdisciplinarity on scientists' research." *Administrative Science Quarterly* 62 (1): 105.

Lerner, Josh, and Ulrike Malmendier. 2013. "With a little help from my (random) friends: Success and failure in post-business school entrepreneurship." *The Review of Financial Studies* 26 (10): 2411–2452.

Levin, Richard C, Alvin K Klevorick, Richard R Nelson, Sidney G Winter, Richard Gilbert, and Zvi Griliches. 1987. "Appropriating the returns from industrial research and development." *Brookings Papers on Economic Activity* 1987 (3): 783–831.

Markman, Gideon D, Donald S Siegel, and Mike Wright. 2008. "Research and technology commercialization." *Journal of Management Studies* 45 (8): 1401–1423.

Marx, Matt, and Aaron Fuegi. 2020. "Reliance on science: Worldwide front-page patent citations to scientific articles." *Strategic Management Journal*.

Murray, Fiona. 2002. "Innovation as co-evolution of scientific and technological networks: exploring tissue engineering." *Research Policy* 31 (8-9): 1389–1403.

Nanda, Ramana, and Jesper B Sørensen. 2010. "Workplace peers and entrepreneurship." *Management Science* 56 (7): 1116–1126.

O'Shea, Rory P, Thomas J Allen, Arnaud Chevalier, and Frank Roche. 2005. "Entrepreneurial orientation, technology transfer and spinoff performance of US universities." *Research Policy* 34 (7): 994–1009.

Rothaermel, Frank T, Shanti D Agung, and Lin Jiang. 2007. "University entrepreneurship: a taxonomy of the literature." *Industrial and Corporate Change* 16 (4): 691–791.

Samila, Sampsa, and Olav Sorenson. 2011. "Venture capital, entrepreneurship, and economic growth." *The Review of Economics and Statistics* 93 (1): 338–349.

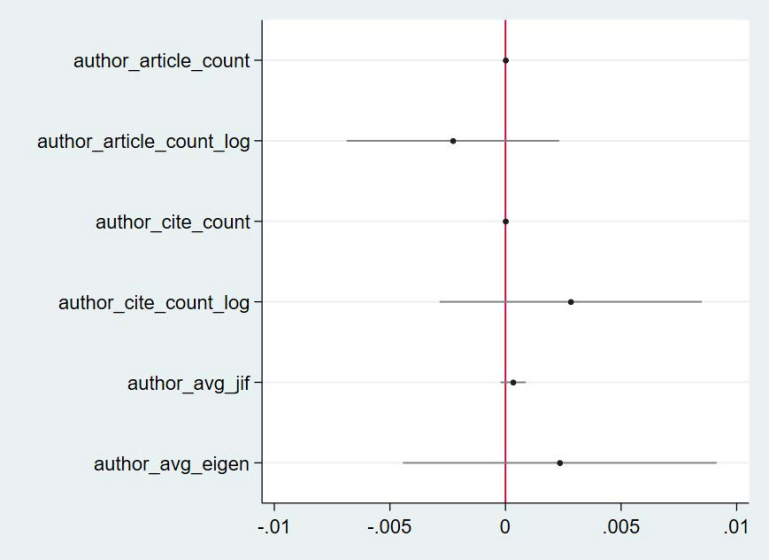
Shane, Scott, and Sankaran Venkataraman. 2000. "The promise of entrepreneurship as a field of research." *Academy of Management Review* 25 (1): 217–226.

Stuart, Toby E, and Waverly W Ding. 2006. "When do scientists become entrepreneurs? The social structural antecedents of commercial activity in the academic life sciences." *American Journal of Sociology* 112 (1): 97–144.

Stuart, Toby E, Ha Hoang, and Ralph C Hybels. 1999. "Interorganizational endorsements and the performance of entrepreneurial ventures." *Administrative Science Quarterly* 44 (2): 315–349.

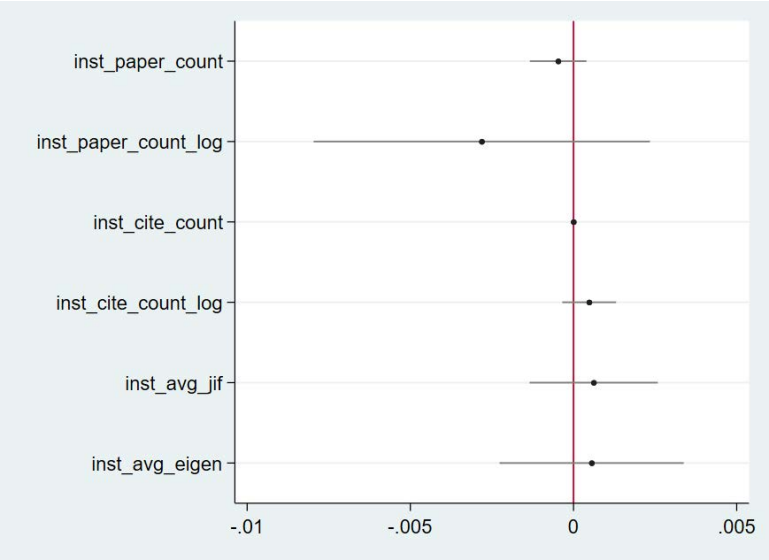
Zucker, Lynne G, Michael R Darby, and Marilynn B Brewer. 1998. "Intellectual Human Capital and the Birth of US Biotechnology Enterprises." *The American Economic Review* 88 (1): 290–306.

Figure 1: Specification map for Author Prestige



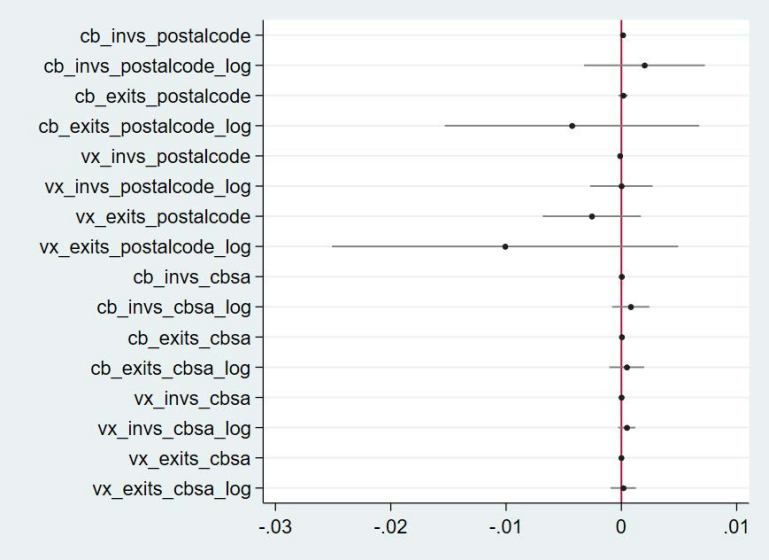
Notes: Coefficient estimates and confidence intervals are based on the final column of Table 3, (author_article_count). Perturbations of author prestige include logged (_log) and unlogged count of papers, a count of citations (_cite_) instead of articles, average JIF for authors on the paper (_jif), and average eigenfactor (_eigen) for authors on the paper. Confidence intervals appear tight for unlogged variables relative to logged variables.

Figure 2: Specification map for Institutional Prestige



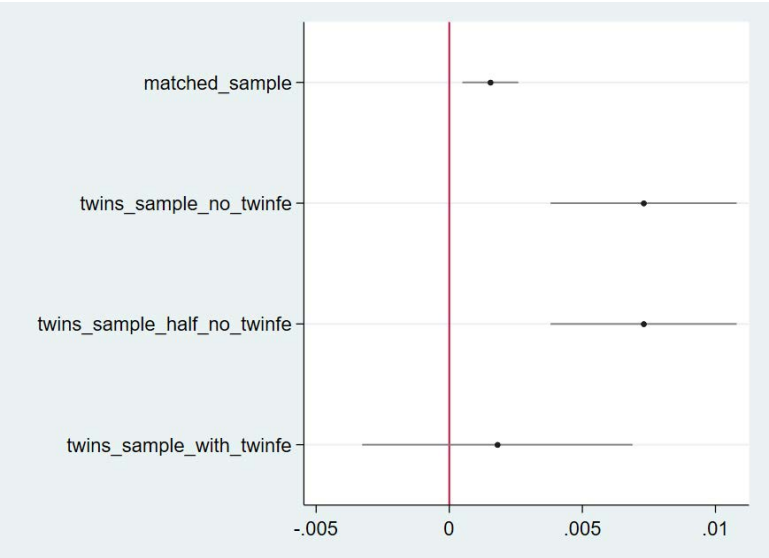
Notes: Coefficient estimates and confidence intervals based on the final column of Table 3, (inst_paper_count). Perturbations of institution prestige include logged (_log) and unlogged count of papers, a count of citations (_cite_) instead of articles, average JIF for authors on the paper (_jif), and average eigenfactor for authors on the paper (_eigen). Confidence intervals appear tight for unlogged variables relative to logged variables.

Figure 3: Specification map for VC investment



Notes: Coefficient estimates and confidence intervals based on the final column of Table 3. Perturbations of VC investment include logged ("_log"), unlogged, and replacing the count of investments per postal code ("_invs_") with a count of liquidity events (i.e., IPOs or acquisitions, "_exits_") per postal code. Additionally, the count of investments via CrunchBase ("cb_") is replaced by that from VenturExpert ("vx_"), and we explore both counts within the postal code ("_postalcode_") and Core-Based Statistical area ("_cbsa_"). Confidence intervals appear tight for unlogged variables relative to logged variables.

Figure 4: VC investment appears to matter until twin FE are applied



Notes: The first row, **matched_sample**, is based on column 8 of Table 2. The second row, **twins_sample_no_twinfe**, is based on the same sample as column 8 of Table 3 but does not include twin fixed effects. The third row, **twins_sample_half_no_twinfe**, resembles the second row but only includes one of the each pair of twin papers. The final row, **twins_sample_with_twinfe**, is based on the final column of Table 3, including twin fixed effects.

Table 1: Difference-of-means tests for twin vs. non-twin papers

Panel A: all twins (23,851) vs. all non-twins (42,051,322)			
	avg. for twins	avg. for non-twins	p<
Author prestige	2.963	1.678	0.00
Institution prestige	6.740	4.899	0.00
Ln same-postalcode investments (CB)	0.752	0.400	0.00
Interdisciplinarity	0.448	0.479	0.00
Prior work w/star commercializer	0.032	0.004	0.00

Panel B: matched twins (19,245) vs. matching non-twins (18,268)			
	avg. for twins	avg. for non-twins	p<
Author prestige	2.966	2.966	0.94
Institution prestige	6.744	6.791	0.01
Ln same-postalcode investments (CB)	0.752	0.752	0.94
Interdisciplinarity	0.448	0.461	0.00
Prior work w/star commercializer	0.032	0.031	0.71

Notes: Table reports difference-of-means tests for 'twin' papers and non-twin papers. Panel A compares twin papers vs. all academic papers from the Web of Science. Panel B compares the subset of twins and Web of Science papers that can be matched using 1:1 Coarsened Exact Matching. Matching variables include author prestige, institution prestige, same-postalcode investments, and whether the authors had prior affiliation with a 'star' commercializer. Non-binary variables were segmented into ten bins each. For computational efficiency, matching was performed on a 10 percent sample of all Web of Science papers.

Table 2: Cross-sectional estimates for startup commercialization of academic science

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Author prestige	0.0006***						0.0000
	(0.0002)						(0.0002)
Institution prestige		-0.0002					-0.0003
		(0.0002)					(0.0002)
Ln same-postalcode investments (CB)			0.0015***				0.0015***
			(0.0005)				(0.0005)
Interdisciplinarity				-0.0007			-0.0001
				(0.0009)			(0.0009)
Star commercializer on paper					0.0814***		0.0431**
					(0.0195)		(0.0206)
Prior work w/star commercializer						0.0479***	0.0404***
						(0.0063)	(0.0064)
Constant	0.0005	0.0033**	0.0010**	0.0024***	0.0017***	0.0009***	0.0017
	(0.0006)	(0.0015)	(0.0005)	(0.0005)	(0.0003)	(0.0002)	(0.0015)
Observations	37618	37618	37618	37618	37618	37618	37618
R^2	0.019	0.019	0.019	0.019	0.030	0.040	0.044
Twin-discovery fixed effects	no	no	no	no	no	no	no

Notes: Sample is as described in Panel B of Table 1, including both the twin discoveries and articles from the full WoS that could be matched closely on observables as shown in Panel B of Table 1. Mean of the dependent variable = 0.003. All models are estimated w/OLS with fixed effects for year, country, and Web of Science subject as well as robust standard errors: *=p<.1; **=p<.05; ***=p<.01. Not shown are controls for number of authors and the count of citations from patents to the scientific articles.

Table 3: Twin-discovery estimates for startup commercialization of academic science

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Author prestige	0.0002						-0.0023
	(0.0014)						(0.0023)
Institution prestige		-0.0020					-0.0028
		(0.0018)					(0.0026)
Ln same-postalcode investments (CB)			0.0017				0.0020
			(0.0027)				(0.0027)
Interdisciplinarity				0.0069*			0.0095**
				(0.0038)			(0.0048)
Star commercializer on paper					0.1307***		0.0959***
					(0.0336)		(0.0352)
Prior work w/star commercializer						0.0582***	0.0406***
						(0.0122)	(0.0123)
Constant	0.0030	0.0052	0.0030	0.0008	0.0034	0.0038	0.0063*
	(0.0034)	(0.0033)	(0.0028)	(0.0033)	(0.0028)	(0.0028)	(0.0033)
Observations	23851	23851	23851	23851	23851	23851	23851
R^2	0.509	0.509	0.509	0.509	0.514	0.514	0.516
Twin-discovery fixed effects	yes	yes	yes	yes	yes	yes	yes

Notes: Observations are articles reporting "twin" academic discoveries, generated as per the procedure described in Section 3. Mean of the dependent variable = 0.009. All models are estimated w/OLS using fixed effects for the twin scientific discovery and robust standard errors: *=p<.1; **=p<.05; ***=p<.01. Not shown are controls for number of authors, the count of citations from patents to the scientific articles, and whether the article was a twin in multiple discoveries.

Table 4: Deeper examination of interdisciplinarity and "star" commercializers

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Author prestige	-0.0023	-0.0014	-0.0027	0.0000	-0.0014	-0.0024	-0.0022	-0.0002
	(0.0023)	(0.0022)	(0.0022)	(0.0021)	(0.0020)	(0.0024)	(0.0024)	(0.0026)
Institution prestige	-0.0028	-0.0031	-0.0009	-0.0039	-0.0033	-0.0029	-0.0029	-0.0033
	(0.0026)	(0.0026)	(0.0028)	(0.0026)	(0.0026)	(0.0026)	(0.0026)	(0.0027)
Ln same-postalcode investments (CB)	0.0020	0.0020	0.0020	0.0020	0.0020	0.0020	0.0020	0.0016
	(0.0027)	(0.0027)	(0.0027)	(0.0027)	(0.0027)	(0.0027)	(0.0027)	(0.0027)
Interdisciplinarity	0.0095**					0.0097**	0.0093*	0.0085*
	(0.0048)					(0.0048)	(0.0048)	(0.0048)
Above-mean interdisciplinarity		0.0039*						
		(0.0022)						
# primary scientific fields among authors			0.0039**					
			(0.0016)					
# authors who publish in one field				0.0002				
				(0.0009)				
Diff (max-mean) author interdisciplinarity					-0.0141**			
					(0.0065)			
Star commercializer on paper	0.0959***	0.0958***	0.0956***	0.0958***	0.0958***			
	(0.0352)	(0.0352)	(0.0352)	(0.0352)	(0.0352)			
Prior work w/star commercializer	0.0406***	0.0406***	0.0405***	0.0407***	0.0411***			
	(0.0123)	(0.0123)	(0.0123)	(0.0123)	(0.0123)			
Star commercializer on paper (50th pctile)						0.1095***		
						(0.0370)		
Prior work w/star commercializer (50th pctile)						0.0347***		
						(0.0114)		
Star commercializer on paper (90th pctile)							0.1084***	
							(0.0354)	
Prior work w/star commercializer (90th pctile)							0.0398***	
							(0.0130)	
Prior work w/star scientist								0.0012
								(0.0023)
Constant	0.0063*	0.0070**	0.0043	0.0073**	0.0204***	0.0064*	0.0062*	0.0036
	(0.0033)	(0.0032)	(0.0036)	(0.0033)	(0.0070)	(0.0033)	(0.0033)	(0.0034)
Observations	23851	23851	23851	23851	23851	23851	23851	23851
R^2	0.516	0.516	0.517	0.516	0.516	0.518	0.517	0.510
Twin-discovery fixed effects	yes	yes	yes	yes	yes	yes	yes	yes

Notes: Observations are articles reporting "twin" academic discoveries, generated as per the procedure described in Section 3. Column (1) repeats column (7) of Table 3. All models estimated w/OLS; robust standard errors: *=p<.1; **=p<.05; ***=p<.01. Mean of the DV= 0.009. Each model includes fixed effects for the twin discovery. Not shown are controls for number of authors, the count of citations from patents to the scientific articles, and whether the article was a twin in multiple discoveries.

Table 5: Robustness and placebo tests for startup commercialization of "twin" discoveries

DV =	commercialization via startup				randomly generated
Sample =	all twins	twins w/variation in DV		twins with multiple adjacency	all twins
	(1)	(2)	(3)	(4)	(5)
Author prestige	-0.0023	-0.1958	-0.0806	-0.0039	-0.0156
	(0.0023)	(0.2399)	(0.1006)	(0.0032)	(0.0110)
Institution prestige	-0.0028	-0.4484	-0.1850	-0.0012	0.0027
	(0.0026)	(0.3616)	(0.1495)	(0.0035)	(0.0147)
Ln same-postalcode investments (CB)	0.0020	0.1123	0.0464	0.0024	-0.0021
	(0.0027)	(0.1042)	(0.0468)	(0.0040)	(0.0082)
Interdisciplinarity	0.0095**	1.1009*	0.4617*	0.0117*	-0.0056
	(0.0048)	(0.6230)	(0.2670)	(0.0065)	(0.0251)
Star commercializer on paper	0.0959***	1.3329*	0.3610*	0.1212***	0.0425
	(0.0352)	(0.8036)	(0.1894)	(0.0446)	(0.0726)
Prior work w/star commercializer	0.0406***	1.2143***	0.5136***	0.0398**	-0.0252
	(0.0123)	(0.4022)	(0.1485)	(0.0166)	(0.0333)
Constant	0.0063*		0.4661***	0.0022	0.5484***
	(0.0033)		(0.1640)	(0.0045)	(0.0158)
Observations	23851	436	436	14721	23851
R^2	0.516		0.140	0.519	0.499
Estimation	OLS	Logit	OLS	OLS	OLS
Twin-discovery fixed effects	yes	yes	yes	yes	yes

Notes: Observations are articles reporting "twin" academic discoveries, generated as per the procedure described in Section 3. Column (1) repeats column (7) of Table 3. Column (2) re-estimates column (1) using logistic regression, which drops twin discoveries whether neither article is commercialized. Column (3) re-estimates (2) using OLS but for the same sample (i.e., where one of the twin articles was commercialized). Column (4) re-estimates (1) for the subset of twins found to have been cited adjacently by multiple papers. For column (5), we generate random values for the dependent variable as a placebo test. Robust standard errors: *=p<.1; **=p<.05; ***=p<.01. Each model includes fixed effects for the twin discovery. Not shown are controls for number of authors, the count of citations from patents to the scientific articles, and whether the article was a twin in multiple discoveries.

Appendices

A Characteristics of twin discoveries

Our 23,851 twin discoveries range from 1973-2015 and are from more than 3,000 academic institutions in 106 countries. Figure A1 shows their temporal distribution. (There may be additional twin discoveries in the distant past, but these are hard to discover because SBIR data are available only since 1983, and patent-to-paper citations are difficult to collect pre-1976 given errors in OCR processing of patent applications. This may also explain why the modal year for a twin discover is somewhat more recent than for the entire Web of Science, 2000 vs. 1997.)

Figure A1: Temporal Distribution of Twin Discoveries

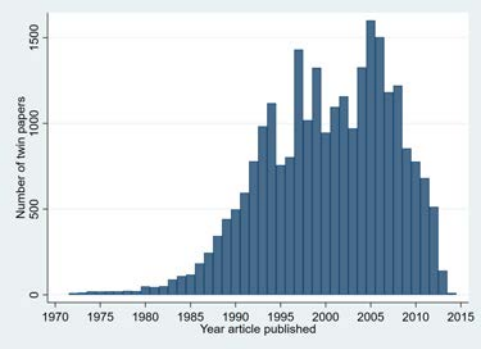


Table A1 shows the distribution of twin discoveries by geography, discipline, and institution. Over half of twin discoveries occur in the U.S., followed by Great Britain, Germany, and Japan. When considering pairs of twin papers, one-third of pairs both occur in the U.S. and 37% of twin papers are in the same country. Panel B details the disciplinary fields of the twin discoveries. The life sciences are responsible for many of the most popular categories of twin discoveries, although Physics is the most popular category. Astronomy Astrophysics is also a frequent source of twin discoveries. Finally, Panel C tabulates the academic institutions with the most twin discoveries.

Table A1: Twin geography, disciplines, and institutions

Panel A		**Panel B**		**Panel C**	
Top 20 countries	%	Top 20 disciplines	%	Top 20 institutions	%
United States	54.1	Physics	6.0	Harvard	3.3
Great Britain	8.1	Cell Biology	5.4	UC San Francisco	1.5
Germany	6.8	Medicine, General & Internal	4.8	Stanford	1.5
Japan	5.3	Genetics & Heredity	4.0	University of Texas	1.4
France	4.5	Immunology	3.7	MIT	1.3
Canada	3.2	Astronomy & Astrophysics	2.9	UC Berkeley	1.3
Netherlands	2.1	Neurosciences	2.9	Yale	1.3
Italy	2.1	Oncology	2.6	Johns Hopkins	1.1
Switzerland	2.0	Developmental Biology	2.0	UC San Diego	1.1
Austria	1.7	Hematology	1.6	Caltech	1.0
Sweden	1.2	Physics, Condensed Matter	1.5	Columbia	0.9
China	1.1	Cardiac & Cardiovascular System	1.5	UCLA	0.9
Israel	0.9	Clinical Neurology	1.3	Cambridge University	0.9
Spain	0.7	Chemistry	1.2	Washington University	0.9
Denmark	0.7	Virology	1.1	University of Washington	0.9
Austria	0.6	Endocrinology & Metabolism	1.0	Tokyo University	0.9
Belgium	0.4	Geochemistry & Geophysics	1.0	University of Pennsylvania	0.8
Finland	0.4	Gastroenterology & Hepatology	0.9	University of Michigan	0.8
South Korea	0.4	Optics	0.9	Oxford University	0.8
Scotland	0.2	Chemistry, Physical	0.8	Rockefeller University	0.8

B Name-overlap and validation for the startup commercialization outcome variable

As described in the main text, our algorithm for determining commercialization relies on overlap between the authors of a paper in the Web of Science and either inventors on a patent or the principal investigators on an SBIR award.

We implement name matching for Web of Science authors vs. SBIR personnel, removing hyphenation and other punctuation. (We examine the first 30 authors on each paper although some papers have more than 30 authors.) Although full names are available for SBIR and patents, many papers only have the authors' surname and initial(s). If both the author and the SBIR awardee have both initials present but these do not match, a score of zero is assigned. Names lacking first initials are ignored. Otherwise, a match score is assigned through a series of steps. First, we determine whether the surnames match exactly or nearly, where "nearly" indicates that both surnames are more than five characters long and fewer than one-fourth of the characters must be changed to convert one to the other (i.e., Levenshtein distance). Moreover, the surnames must start with the same letter (e.g., "Rogers" and "Bogers" are not matched). Two names are treated as a preliminary match if the surname meets these criteria and the first initials also match. We want to avoid the situation where the author "J Smith" is assumed to be the same as the SBIR awardee "Jesse Smith", so we score surnames according to their inverse frequency of appearance in the Web of Science. For instance, surname Smith would be downscaled to near-zero as it is among the most common author names. Surnames that comprise less than 0.007% of all authors (i.e., 2nd percentile) are not downscaled. If only two authors match between the paper and SBIR grant, and both of them represent more than 0.005% of all authors, we conclude that there is no match. Regardless of surname, matches are considered exact if both first and second initials are present for both names and they both match. A similar algorithm is implemented for computing overlap between authors of articles and inventors on patents.

To evaluate whether our algorithm truly captures instances of startup commercialization, we examine a random sample of both types of potential examples of commercialization to seek direct confirmation of our algorithmic approach. Table B1 shows five of the 20 examples of paper-patent pairs we researched, and Table B2 shows five of the 20 examples of SBIR grants. We start by randomly selecting 20 scientific papers drawn from each route of identifying commercialization. For each of these papers, we retrieve the underlying scientific article via Google Scholar searches and record the authors. For Table B1, we retrieve the associated patent from our algorithmic approach described in the main text via Google Patents (patents.google.com). We record the patent title, inventors, and assignee. For Table B2, we retrieve the associated SBIR grants to the focal companies via sbir.gov and record the grant title, funding agency and amount, and the listed principal investigator/business contact. To verify the linkages in both panels between scientific paper and commercialization activity, we conduct web searches in the following manner: we find the overlapping names between paper author and patent inventor or SBIR contact—those are shown in bold in the table. We search the web for the union of the overlapped name(s) and the new venture entity. The final column in both tables provide web links (all accessed in January 2019) providing confirmation of commercialization activity in all ten instances (in the broader sample, we verified 39 out of 40 overall cases).

One interesting case is the second entry in Table B1. We initially had difficulty finding confirmation, but then found that one of the author/inventors, Larry Gold, had founded a company, NeXagen to commercialize his technology, changed the name of the company, and subsequently sold that company to Gilead Sciences. The patent was subsequently reassigned to Gilead Sciences, which is why initially we thought we had failed to find a linkage.

Table B1: Random sample of five patent-paper-pair instances of startup commercialization

Paper title	Journal / Year	Authors	Institution	Patent	Inventors	Patent assignee	Linkages
RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions	*PNAS* / 2011	Wiedenheft, B; van Duijn, E; Bultema, JB; Waghmare, SP; Dickman, M; Zhou, KH; Barendregt, A; Westphal, W; **Doudna, JA**	Univ Calif Berkeley	Compositions and methods of nucleic acid-targeting nucleic acids (9260752)	Andrew Paul May; **Rachel E. Haurwitz**; **Jennifer A. Doudna**; James M. Berger; Matthew Merrill Carter; Paul Donohoue	Caribou Biosciences, Inc.	**Doudna** is on Caribou's SAB; **Haurwitz** is Caribou's CEO and on the firm's BoD. Source: https://caribou bio.com/about-us
Systematic evolution of ligands by exponential enrichment - RNA ligans to bacteriophage-T4 DNA-polymerase	*Science* / 1990	Tuerk, C; **Gold, L**	Univ Colorado	Systematic evolution of ligands by exponential enrichment: tissue selex (6613526)	Joseph S. Heilig; **Larry Gold**	Gilead Sciences, Inc.	**Gold** is a founder of NeXagen, which became NeXstar Pharmaceuticas. That organization merged with Gilead Sciences in 1999. Source: https://somalogic.com/about-us/leadership/larry-gold-2/
Phase selection of microcrystalline GaN synthesized in supercritical ammonia	Journal of Crystal Growth / 2006	**Hashimoto, T**; Fujito, K; Sharma, R; **Letts, ER**; Fini, PT; Speck, JS; Nakamura, S	Univ Calif Santa Barbara	Method for producing group III-nitride wafers and group III-nitride wafers (9803293)	**Tadao Hashimoto**; Edward Letts; Masanori Ikari	SixPoint Materials Inc	**Hashimoto** is CEO/CTO of SixPoint; **Letts** is VP of Technology of the firm. Source: http://www.spmaterials.com/team.htm
Preoperative Diagnosis of Benign Thyroid Nodules with Indeterminate Cytology	*NEJM* / 2012	Alexander, EK; **Kennedy, GC;** Baloch, ZW; Cibas, ES; Friedman, L; Lanman, RB; Mandel, SJ ;Yener, N; Kloos, RT; LiVolsi, VA; Lanman, RB; Steward, DL; Friedman, L; Kloos, RT; Wilde, JI; Raab, SS; Haugen, BR; Steward, DL; Zeiger, MA; Haugen, BR	Brigham & Womens Hospital	Algorithms for disease diagnostics (9495515)	**Giulia C. Kennedy**; Darya I. Chudova; Eric T. Wang; Jonathan I. Wilde	Veracyte Inc	**Kennedy** is Chief Scientific and Medical Officer of Veracyte. https://www.veracyte.com/who-we-are/leadership/executive-team. Wilde was a director and VP of Discovery Research at Veracyte. https://uk.linkedin.com/in/jonathanwilde650
Human retinoblastoma susceptibility gene - cloning, identification, and sequence	*Science* / 1987	**Lee, WH**; Bookstein, R; Hong, F; Young, LJ; Shew, JY; Lee, EYHP	Univ Calif San Diego	Therapeutic use of the retinoblastoma susceptibility gene product (5851991)	**Wen-Hwa Lee**; Eva Y-H.P. Lee; David W. Goodrich; H. Michael Shepard; Nan Ping Wang; Duane Johnson	University of California; Canji Inc	**Wen-Hwa Lee** was Chair of the Scientific Advisory Board of Canji, Inc. http://rcndd.cmu.edu.tw/sites/default/files/WHL-CV.pdf. Canji was "formed to commercialize suppressor oncogene technology developed by Dr. Wen-Hwa Lee of the University of California at San Diego. Canji, Inc. operates as a subsidiary of Merck & Co." https://www.bloomberg.com/research/stocks/private/snapshot.asp?privcapid=26032.

Table B2: Random sample of five SBIR instances of startup commercialization

Paper title	Journal / Year	Authors	Institution	SBIR Company	SBIR Grant(s)	SBIR PIs	Linkages
The outer mitochondrial membrane protein mitoNEET contains a novel redox-active 2Fe-2S cluster	Journal of Biological Chemistry / 2007	Wiley, SE; Paddock, ML; Abresch, EC; Gross, L; van der Geer, P; Nechushtai, R; Murphy, AN; Jennings, PA; Dixon, JE	Univ Calif San Diego	Mitokor, Inc.	"Mitochondrial Functional Proteomics" (2005 for $100,000 from the Department of Defense); "Osteoarthritis/Chondrocalcinosis: Mitochondrial Therapy" ($106,745 from the Department of Health and Human Services (HHS))	Eoin Fahy; Anne Murphy	Murphy was Director of Mitochondrial Biology at MitoKor: https://www.researchgate.net/profile/Anne_Murphy/2
Scattering theory derivation of a 3D acoustic cloaking shell	Physical Review Letters / 2008	Cummer, SA; Popa, B; Schurig, D; Smith DR; Pendry, J; Rahm, M; Starr A	Duke Univ	SensorMetrix, Inc.	"Development of Acoustic Metamaterial Applications" ($750,813 from the Dept of Defense (Navy))	Anthony Starr	Dr. Anthony Starr is the founder, president & CEO of SensorMetrix. http://www.sensormetrix.com/key-personnel.html
Global sequencing of proteolytic cleavage sites in apoptosis by specific labeling of protein N termini	Cell / 2008	Mahrus, S; Trinidad, JC; Barkan, DT; Sali, A; Burlingame, AL; Wells, JA	Univ Calif San Francisco	Sunesis Pharmaceuticals, Inc.	"Development of Conformation Specific Kinase Inhibitors" (HHS for $1.5M)	James A. Wells	Wells is founder of Sunesis Pharmaceuticals. https://www.crunchbase.com/person/jim-wells#section-jobs and https://www.bloomberg.com/research/stocks/private/person.asp?personId=467474&privcapId=3768647&previousCapId=177932577&previousTitle=REZOLUTE%20INC
Curved plasma channel generation using ultraintense airy beams	Science / 2009	Polynkin, P; Kolesik, M; Moloney, JV; Siviloglou, GA; Christodoulides, DN	Univ Arizona	Nonlinear Control Strategies, Inc.	"High Power, Room Temperature 2.4- 4 micron Mid-IR Semiconductor Laser Optimization" (Department of Defense (Air Force) for $99,995 and $746,925	Jerome V Moloney	Moloney is President and corporate head of Nonlinear Control Strategies. http://www.nlcstr.com/contact.htm
Whole-genome sequencing identifies recurrent somatic NOTCH2 mutations in splenic marginal zone lymphoma	Journal of Experimental Medicine / 2012	Kiel, MJ; Velusamy, T; Betz, BL; Zhao, L; Weigelin, HG; Chiang, MY; Huebner-Chan, DR; Bailey, NG; Medeiros, LJ; Bailey, NG; Elenitoba-Johnson, KSJ	Univ Michigan	Genomenon, Inc.	"Commercial Software Using High throughput Computational Techniques to Improve Genome Analysis" (HHS- National Institutes of Health, $972,083)	Mark Kiel	Kiel is a co-founder of Genomenon and Chief Science Officer. https://www.genomenon.com/about/; https://www.crunchbase.com/organization/genomenon

C Characteristics of 'star' commercializers

Appendix Table C1 provides additional information on the nature of star entrepreneurial commercializers. Only 0.4% of the more than 73 million authors in the Web of Science have had one of their discoveries commercialized by a startup. The vast majority of authors whose discoveries are commercialized by startups do so only once (mean = 1.26). Overall, less than 0.01% of all authors are ever "stars" in this respect.

Panel A of Appendix Table C1 compares stars with all other authors in the Web of Science. Perhaps unsurprisingly, stars have many more articles and citations per article, and they have been publishing longer than non-stars. Panel B details the most popular fields among stars, using 251 fields from the Web of Science. Biochemistry & Molecular Biology is the most frequent field for entrepreneurial commercialization (13.2% of all stars work primarily in this field), followed by Chemistry, Electrical & Electronic Engineering, Immunology, and Applied Physics. Panel C shows the frequency of "star" involvement in commercialized discoveries by industry and time period.

Table C1: Descriptive statistics for 'star' entrepreneurial commercializers

Panel A: Star commercializers vs. all other authors (N: 7,164 vs. 73,923,279)

	avg. non-star	avg. star	stderr	$p <$
lifetime num articles	1.639	13.708	0.040	0.000
average citations per paper	13.179	30.961	0.555	0.000
num years publishing	0.899	7.432	0.035	0.000

Panel B: Most popular fields for star commercializers

Field of study	Percent of stars in that field
Biochemistry and molecular biology	13.2
Chemistry, multidisciplinary	6.5
Engineering, electrical and electronic	5.1
Immunology	4.5
Physics, applied	4.2
Oncology	3.9
Multidisciplinary Sciences	3.6
Chemistry, medicinal	3.6
Cardiac and cardiovascular systems	3.3
Endocrinology and metabolism	2.9

Panel C: Prevalence of star commercializers among commercialized discoveries

	pre-2000	2000 and later
biotech	0.27	0.22
non-biotech life sciences	0.00	0.03
non life-sciences	0.06	0.05

D Geographic, industry, and temporal heterogeneity

In Table D2 we attempt to understand whether the results from Table 3 are driven by geography, scientific field, or time periods, with the caveat that splitting the sample along these dimensions may yield noisier estimation due to lack of statistical power. Columns (1-3) split the sample of twins into three groups: (1) both twins are in the U.S.; (2) neither of the twins in the scientific discovery is in the U.S.; (3) the scientific discovery contains a twin from the U.S. as well as a twin from outside the U.S. The majority of twins are "mixed"—i.e., in the third category. Twins are somewhat more often commercialized when the discovery teams are interdisciplinary, but only as long as one of the twins is in the U.S. The effect of having a "star" commercializer on the paper is also strongest within U.S.-only twins. Prior collaboration with a star commercializer appears to play a role primarily for twins where both articles are from outside the U.S..

Table D2: Heterogeneity in commercialization by geography, industry, and time period

	both twins in U.S.	neither twin in U.S.	only one in U.S.	biotech	non-biotech life sciences	not life sciences	pre-2000	post-1999
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Author prestige	-0.0050	-0.0014	-0.0015	0.0004	-0.0047	0.0018	-0.0035	-0.0004
	(0.0056)	(0.0017)	(0.0037)	(0.0047)	(0.0053)	(0.0038)	(0.0023)	(0.0034)
Institution prestige	-0.0062	-0.0017	-0.0008	-0.0020	-0.0054	-0.0054	0.0034	-0.0072*
	(0.0055)	(0.0017)	(0.0046)	(0.0062)	(0.0079)	(0.0042)	(0.0025)	(0.0041)
Ln same-postalcode investments (CB)	-0.0007	0.0092	0.0034	0.0047	0.0029	0.0027	0.0177	0.0013
	(0.0046)	(0.0062)	(0.0035)	(0.0068)	(0.0057)	(0.0039)	(0.0130)	(0.0027)
Interdisciplinarity	0.0186*	-0.0046	0.0123*	-0.0038	0.0266**	0.0024	0.0129**	0.0055
	(0.0103)	(0.0045)	(0.0074)	(0.0082)	(0.0112)	(0.0069)	(0.0056)	(0.0075)
Star commercializer on paper	0.1440**	0.0041	0.0731	0.2318**	0.1656***	-0.0282	0.2631***	0.0350
	(0.0587)	(0.0539)	(0.0505)	(0.0908)	(0.0602)	(0.0758)	(0.0781)	(0.0368)
Prior work w/star commercializer	0.0372*	0.0519**	0.0377*	-0.0111	0.0688***	0.0343	0.0238	0.0457***
	(0.0210)	(0.0215)	(0.0202)	(0.0209)	(0.0224)	(0.0239)	(0.0210)	(0.0146)
Constant	0.0135*	0.0068**	0.0004	0.0017	0.0033	0.0091**	-0.0112**	0.0150***
	(0.0072)	(0.0033)	(0.0053)	(0.0057)	(0.0085)	(0.0046)	(0.0046)	(0.0044)
Observations	7910	5969	9972	5139	6189	6778	9389	14090
R^2	0.528	0.510	0.503	0.519	0.514	0.523	0.549	0.510
Twin-discovery fixed effects	yes	yes	yes	yes	yes	yes	yes	yes

Notes: Observations are articles reporting "twin" academic discoveries, generated as per the procedure described in Section 3. Columns (1-3) explore geographic variation depending on whether both, neither, or just one of the twins is in the U.S. Columns (4-6) subsample the twins data by whether they are from biotechnology, non-biotech life sciences, or outside the life sciences, respectively. Finally, columns (7) & (8) split the sample by time period. All models estimated w/OLS; robust standard errors: *=p<.1; **=p<.05; ***=p<.01. Mean of the DV= 0.009. Each model includes fixed effects for the twin discovery. Not shown are controls for number of authors, the count of citations from patents to the scientific articles, and whether the article was a twin in multiple discoveries.

Regarding scientific field, much of our collective empirical knowledge on commercialization draws on the field of biotechnology, perhaps due to data availability reasons, as previously discussed. In column (4), having a star commercializer on the paper predicts commercialization of biotechnology discoveries, but neither interdisciplinarity nor prior association with a star does. Of course, the life sciences are not limited to biotechnology; in column (5) we analyze non-biotechnology papers in life sciences and find that interdisciplinarity plays a key role. Prior association with a star commercializer plays a role as well. Finally, in column (6) we analyze non-life-sciences papers, of which there are more than either of the other categories. We find no statistically significant covariates.

Columns (7) and (8) attempt to situate these results temporally, splitting the sample into papers published before and after the year 2000. It appears that interdisciplinarity played a significant role only for papers published in the 20th century. Moreover, we observe a shift from reliance on having a star among the authors of the paper earlier (column 7) to prior association with a star (column 8).