Does Science Advance One Funeral at a Time?

Pierre Azoulay
MIT and NBER
Sloan School of Management
100 Main Steet – E62-487
Cambridge, MA 02142
USA

Christian Fons-Rosen
Universitat Pompeu Fabra and CEPR
Barcelona GSE
Carrer Ramon Trias Fargas, 25-27
08005 Barcelona
Spain

Joshua S. Graff Zivin
UCSD and NBER
School of Global Policy & Strategy
9500 Gilman Drive
La Jolla, CA 92093
USA

June 19, 2018

Abstract

We study the extent to which eminent scientists shape the vitality of their areas of scientific inquiry by examining entry rates into the subfields of 452 academic life scientists who pass away prematurely. Consistent with previous research, the flow of articles by collaborators into affected fields decreases precipitously after the death of a star scientist. In contrast, we find that the flow of articles by non-collaborators increases by 8.6% on average. These additional contributions are disproportionately likely to be highly cited. They are also more likely to be authored by scientists who were not previously active in the deceased superstar's field. Intellectual, social, and resource barriers all impede entry, with outsiders only entering subfields that offer a less hostile landscape for the support and acceptance of "foreign" ideas. Overall, our results suggest that once in control of the commanding heights of their fields, star scientists tend to hold on to their exalted position a bit too long.

Keywords: economics of science, scientific fields, superstars, invisible college, cumulative knowledge production.

*Address all correspondence to pazoulay@mit.edu. Azoulay and Graff Zivin acknowledge the financial support of the National Science Foundation through its SciSIP Program (Award SBE-1460344). Christian Fons-Rosen acknowledges financial support from the Spanish Ministry of Economy and Competitiveness through a grant (ECO-2014-55555-P) and through the Severo Ochoa Programme for Centres of Excellence in R&D (SEV-2015-0563). Mikka Rokkanen provided additional research assistance. The project would not have been possible without Andrew Stellman's extraordinary programming skills (www.stellman-greene.com). We thank Heidi Williams, Xavier Jaravel, Danielle Li, Sameer Srivastava, Scott Stern, Bruce Weinberg, and seminar audiences at the NBER, UC Berkeley, National University of Singapore, and Stanford University for useful discussions. The usual disclaimer applies.

"A new scientific truth does not triumph by convincing its opponents and making them see the light, but rather because its opponents eventually die, and a new generation grows up that is familiar with it."

MAX PLANCK
Scientific Autobiography and Other Papers

1 Introduction

Whether manna from heaven or the result of the purposeful application of research and development, technological advances play a foundational role in all modern theories of economic growth (Solow 1957, Romer 1990, Aghion and Howitt 1992). Only in the latter part of the nineteenth century, however, did technological progress start to systematically build upon scientific foundations (Mokyr 1990, 2002). Economists—in contrast to philosophers, historians, and sociologists (Kuhn 1962, Shapin 1996, Merton 1973)—have devoted surprisingly little effort to understanding the processes and institutions that shape the evolution of science.[1] How do researchers identify problems worthy of study and choose among potential approaches to investigate them?

Presumably these choices are driven by a quest for recognition and scientific glory, but the view that scientific advances are the result of a pure competition of ideas—one where the highest quality insights inevitably emerge as victorious—has long been considered a Panglossian but useful foil (Kuhn 1962; Akerlof and Michaillat 2017). Indeed, the provocative quote from Max Planck in the epigraph of this paper underscores that even the most celebrated scientist of his era understood that the pragmatic success of a scientific theory does not entirely determine how quickly it gains adherents, or its longevity.

Can the idiosyncratic stances of individual scientists do much to alter, or at least delay, the course of scientific advance? Perhaps for the sort of scientific revolutions that Planck—the pioneer of quantum mechanics—likely had in mind, but the proposition that established scientists are slower than novices in accepting paradigm-shifting ideas has received little empirical support whenever it has been put to the test (Hull et al. 1978; Gorham 1991; Levin et al. 1995). Paradigm shifts are exceedingly rare, however, and their very nature suggests

[1] A notable exception is the theoretical model of scientific revolutions developed by Bramoullé and Saint-Paul (2010).

that once they emerge, it is exceedingly costly to resist or ignore them. In contrast, "normal" scientific advance—the regular work of scientists theorizing, observing, and experimenting within a settled paradigm or explanatory framework—may be more susceptible to political jousting. The absence of new self-evident and far reaching truths means that scientists must compete in a crowded intellectual landscape, sometime savagely, for the supremacy of their ideas (Bourdieu 1975).

In this paper, we use a difference-in-differences setup to test "Planck's Principle" in the context of academic biomedical research, an enormous domain which has been the province of normal scientific change ever since the "central dogma" of molecular biology (Crick 1970) emerged as a unifying description of the information flow in biological systems. Specifically, we examine how the premature death of 452 eminent scientists alter the vitality (measured by publication rates and funding flows) of subfields in which they actively published in the years immediately preceding their passing, compared to matched control subfields. In contrast with prior work that focused on collaborators (Azoulay et al. 2010; Oettl 2012; Jaravel et al. 2018; Mohnen 2018), our work leverages new tools to define scientific subfields which allows us to expand our focus to the response by scientists who may have similar intellectual interests with the deceased stars without ever collaborating with them.

To our surprise, it is not competitors from within a subfield that assume the mantle of leadership, but rather entrants from other fields that step in to fill the void created by a star's absence. Importantly, this surge in contributions from outsiders draws upon a different scientific corpus and is disproportionately likely to be highly cited. Thus, consistent with the contention by Planck, the loss of a luminary provides an opportunity for fields to evolve in novel directions that advance the scientific frontier. The rest of the manuscript is dedicated to elucidating the mechanisms responsible for this phenomenon.

It does not appear to be the case that stars use their influence over financial or editorial resources to block entry into their fields, but rather that the very prospect of challenging a luminary in the field serves as a deterrent for entry by outsiders. Indeed, most of the entry we see occurs in those fields that lost a star who was especially accomplished. Even in those fields that have lost a particularly bright star, entry can still be regulated by key collaborators left behind. We find suggestive evidence that this is true in fields that have coalesced around a narrow set of techniques or ideas or where collaboration networks are particularly tight-knit. We also find that entry is more anemic when key collaborators of the star are in positions that allow them to limit access to funding or publication outlets to those outside the club that once nucleated around the star.

To be clear, we are not arguing that stars are a net negative for scientific progress. Indeed, given the outsized accomplishments of the eminent scientists in our sample, this seems quite unlikely. Rather, our results suggest that, once in control of the commanding heights of their fields, star scientists tend to hold on to their exalted position—and to the power that comes with it—a bit too long. Moreover, the entrants who boost activity into the subfields formerly occupied by the deceased star are disproportionately likely to be future stars, providing evidence that the outsiders of today can often turn into the stars of tomorrow. This "circle of scientific life" underscores the importance of caution when drawing welfare implications about the existence of stars writ large.

To our knowledge, this manuscript is the first to examine the dynamics of scientific evolution using the standard empirical tools of applied microeconomics.[2] We conceptualize the death of eminent scientists as shocks to the structure of the intellectual neighborhoods in which they worked several years prior to their death, and implement a procedure to delineate the boundaries of these neighborhoods in a way that is scalable, transparent, and does not rely on *ad hoc* human judgment. The construction of our dataset relies heavily on the *PubMed Related Citations Algorithm* [PMRA], which groups scientific articles into subfields based on their intellectual content using abstract words, title words, and very detailed keywords drawn from a controlled vocabulary thesaurus curated by the National Library of Medicine. As such, we are able to delineate circumscribed areas of scientific inquiry whose boundaries are not defined by shared training, collaboration, or citation relationships.

In addition to providing evidence regarding a central question for scholars studying the scientific process, our paper is among the very few economic studies that attend to the ways in which scientists position themselves in intellectual space (cf. Borjas and Doran [2015a, 2015b] and Myers [2018] for other notable examples). As such, our work can be understood as integrating the traditional concerns of economists—understanding how incentives and

[2]Considerable work outside of economics has examined the evolution of scientific fields through network and community detection techniques (e.g., Rosvall & Bergstrom 2008; Börner, Chen, and Boyack 2003; cf. Fortunato and Hric (2016) for a review of this fast evolving research area). These approaches rely on collaboration or citation links to define the vertices of the knowledge network used to partition a scientific space into subfields. While social scientists have utilized these techniques to explain a wide range of phenomena (e.g., Foster, Rzhetsky, and Evans 2015), these approaches are less well-suited to our setting where citation and collaboration are among the primary outcomes of interest.

institutions influence the rate of knowledge production or diffusion—with those of cognate disciplines such as sociology and philosophy, who have traditionally taken the *direction* of scientific change as the central problem to be explained.

The rest of the paper proceeds as follows. In the next section, we examine the institutional context and lay out our broad empirical strategy. In section 3, we then turn to data, methods and descriptive statistics. We report the results in section 4. Section 5 concludes by outlining the implications of our findings for future work.

2 Institutional Context and Empirical Design

Our empirical analyses are centered on the academic life sciences. The merits of this focus are several fold. First, the field has been an important source of scientific discovery over the past half century. Many modern medical therapies can trace their origins to research conducted in academic laboratories (Sampat and Lichtenberg 2011; Azoulay, Li, and Sampat 2017). These discoveries, in turn, have generated enormous health and welfare gains for economies around the world.

Second, the life science research workforce is exceedingly large and specialized. The Faculty Roster of the Association of American Medical Colleges lists more than 200,000 faculty members employed in U.S. medical schools and academic medical centers in 2015.[3] Moreover, scientific discoveries over the past half-century have greatly expanded the knowledge frontier, necessitating increasing specialization by researchers and a greater role for collaboration (Jones 2009). If knowledge and techniques remain at least partially tacit long after their initial discovery, tightly-knit research teams may be able to effectively control entry into intellectual domains. The size and maturity of this sector, including its extensive variety of narrowly-defined subfields, makes it an ideal candidate for an inquiry into the determinants of the direction of scientific effort in general, and how it is influenced by elite scientists in particular.

Third, the academic research setting also offers the practical benefits of an extensive paper trail of research inputs, outputs, and collaboration histories. On the input side, reliance of researchers on one agency for the majority of their funding raises the possibility

[3] This figure excludes life science academics employed in graduate schools of arts and science or other non-medical school settings such as MIT, Rockefeller University, The Salk Institute, UC Berkeley, the intramural campuses of NIH, etc.

that financial gatekeeping by elite scientists could be used to regulate entry into scientific fields. Data on NIH funding at the individual level, as well as membership in "study sections" (the peer-review panels that evaluate the scientific merits of grant applications) will allow us to examine such concerns directly. Most importantly for our study, the principal output of researchers—publications—are all indexed by a controlled vocabulary of keywords managed by the National Library of Medicine. This provides the raw material that helps delineate scientific subfields without appealing to citation linkages or collaborative relationships (the specifics of this process are described in detail in Section 3.2 and Appendix C).

These many virtues, however, may come at the expense of generalizability. While the life sciences span a wide range of research styles—from small-team data-driven epidemiology, to medium-size laboratories under the helm of a single principal investigator, to large-scale multi-institution clinical trials—most biomedical researchers cluster topically and socially in small, quasi-independent subfields. This broad domain seldom features exceedingly small research teams (as in pure mathematics) or "big science" efforts where capital needs are so extensive and specialized as to fully consolidate the field into a single or a handful of large authorship teams (as in high-energy particle physics, e.g., Aad et al. 2015). As such, one should refrain from applying our findings to other fields of science where the structure of collaborative efforts and the degree of intellectual clustering are likely to generate different patterns of succession, compared to those observed in the life sciences.

Accounts by practicing scientists indicate that collaboration plays a large role in both the creation and diffusion of new ideas (Reese 2004), and historians of science have long debated the role of controversies and competition in shaping the direction of scientific progress and the process through which new subfields within the same broad scientific paradigm are born and grow over time (Hull 1988; Morange 1998; Shwed and Bearman 2010). Our study presents a unique opportunity to test some of their insights in a way that is more systematic and can yield generalizable insights on the dynamics of field evolution.

3 Empirical Design, Data, and Descriptive Statistics

Below, we provide a detailed description of the process through which the matched scientist/subfield dataset used in the econometric analysis was assembled. We begin by describing the criteria used to select our sample of superstar academics, with a particular focus on "extinction events"; the set of subfields in which these scientists were active prior to their death and the procedure followed to delineate their boundaries. Finally, we discuss the matching procedure implemented to identify control subfields associated with eminent scientists who did not pass away but are otherwise similar to our treatment group.

3.1 Superstar sample

Our basic approach is to rely on the death of "superstar" scientists as a lever to estimate the extent to which the production of knowledge in the fields in which they were active changes after their passing. The study's focus on the scientific elite can be justified both on substantive and pragmatic grounds. The distribution of publications, funding, and citations at the individual level is extremely skewed (Lotka 1926; de Solla Price 1963) and only a tiny minority of scientists contribute, through their published research, to the advancement of science (Cole and Cole 1972). Stars also leave behind a corpus of work and colleagues with a stake in the preservation of their legacy, making it possible to trace back their careers, from humble beginnings to wide recognition and acclaim.

The elite academic life scientist sample includes 12,935 individuals, which corresponds to roughly 5% of the entire relevant labor market. In our framework, a scientist is deemed elite if they satisfy at least one of the following criteria for cumulative scientific achievement: (1) highly funded scientists; (2) highly cited scientists; (3) top patenters; and (4) members of the National Academy of Sciences or of (5) the National Academy of Medicine. Since these criteria are based on extraordinary achievement over an entire scientific career, we augment this sample using additional criteria to capture individuals who show great promise at the early and middle stages of their scientific careers (so-called "shooting stars"). These include: (6) NIH MERIT awardees; (7) Howard Hughes Medical Investigators; and (8) early career prize winners. Appendix A provides additional details regarding these metrics of "superstardom" and explores the sensitivity of our core set of results to the type of scientists ("cumulative stars" vs. "shooting stars") included in the sample.

For each scientist, we reconstruct their career from the time they obtained their first position as independent investigators (typically after a postdoctoral fellowship) until 2006. Our dataset includes employment history, degree held, date of degree, gender, and department affiliations as well as complete list of publications, patents and NIH funding obtained in each year by each scientist.[4]

The 452 scientists who pass away prematurely, and who are the particular focus of this paper, constitute a subset of this larger pool of 12,935. To be included in our sample, their deaths must intervene between 1975 and 2003 (this allows us to observe at least three years' worth of scientific output for every subfield after the death of a superstar scientist). Although we do not impose any age cutoff, the median and mean age at death is 61 with 85% of these scientists having passed away before the age of 70 (we explore the sensitivity of our results to the age at death in Appendix E). We also require evidence, in the form of published articles and/or NIH grants, that these scholars were still in a scientifically active phase of their career in the period just preceding their death (this is the narrow sense in which we deem their deaths to have occurred prematurely).

Within this sample, 229 (51%) of these scientists pass away after a protracted illness, whereas 185 (41%) die suddenly and unexpectedly. We were unable to ascertain the particular circumstances of 37 (8.20%) death events.[5] Table 1 provides descriptive statistics for the deceased superstar sample. The median star received her degree in 1957 and died at the age of 61. 40% of the stars hold an MD degree (as opposed to a PhD or MD/PhD), and 90% of them are male. On the output side, the stars each received an average of roughly 16.6 million dollars in NIH grants, and published 138 papers that garnered 8,341 citations over the course of their careers (as of 2015).

3.2 Delineating Research Fields

The source of the publication data is *PubMed*, an online resource from the National Library of Medicine that provides fast, free, and reliable access to the biomedical research literature. *PubMed* indexes more than 40,000 journals within the life sciences.

To delineate the boundaries of the research fields in which each deceased star was active, we develop an approach based on topic similarity between each article where the star

[4] Appendix B details the steps taken to ensure that the list of publications is complete and accurate, even in the case of stars with frequent last names. Though we apply the term of "star" or "superstar" to the entire group, there is substantial heterogeneity in intellectual stature within the sample (see Table 1).

[5] Table A3 in Appendix A provides the full list of deceased superstars, together with their year of birth and death, cause of death, institutional affiliation at the time of their passing, and a short description of their research expertise.

appeared as a last author in a window of five years prior to her death, and the rest of the scientific literature.[6] Specifically, we use the *PubMed Related Citations Algorithm* (Lin and Wilbur 2007) which relies heavily on Medical Subject Headings (MeSH), but not in any way on citation or collaboration linkages.

MeSH terms constitute a controlled vocabulary maintained by the National Library of Medicine that provides a very fine-grained partition of the intellectual space spanned by the biomedical research literature. Importantly for our purposes, MeSH keywords are assigned to each publication by professional indexers who focus solely on their scientific content. That said, the *PubMed* Related Citations Algorithm (hereafter PMRA) also uses title and abstract words as inputs, which are selected by the authors, and may reflect their aspirations. While this raises the possibility that our subfield definitions are not impervious to social influences, it does offer one advantage, namely that our subfield boundaries can quickly reflect the emergence of new terms whose inclusion in the official MeSH thesaurus will occur with some lag.[7] Regardless, as will become clear in the next section, our difference-in-differences design alleviates the concern that idiosyncratic features of PMRA might affect our conclusions, since these would influence treatment and control subfields in a symmetric fashion.

We then use the "Related Articles" function in *PubMed* to harvest journal articles that are intellectually proximate to the star scientists' own papers in the last five years of her life.[8] Appendix C describes the algorithm in more detail and performs extensive robustness checks. In particular, we verify that the cutoff rules used by PMRA to generate a set of intellectual neighbors for a given source article do not induce treated subfields to exhibit idiosyncratic truncation patterns—from above or from below—compared to control subfields. Using a tunable version of PMRA, we also assess the robustness of our core results to manipulations

[6]A robust social norm in the life sciences systematically assigns last authorship to the principal investigator, first authorship to the junior author who was responsible for the conduct of the investigation, and apportions the remaining credit to authors in the middle of the authorship list, generally as a decreasing function of the distance from the extremities (Zuckerman 1968; Nagaoka and Owan 2014). Only in the case of last authorship can we unambiguously associate the star with a subfield.

[7]Importantly, defining subfields as isomorphic to the set of articles related (in a PMRA-sense) to a source article does not imply a fixed number of articles per subfield. On the contrary, PMRA-generated subfields can be of arbitrary large size. In Appendix C, we document the variation in subfield size and explore the sensitivity of our results to alternate subfield definitions, including those that exclude potentially endogenous intellectual linkages.

[8]To facilitate the harvesting of *PubMed*-related records on a large scale, we have developed an open-source software tool that queries *PubMed* and PMRA and stores the retrieved data in a MySQL database. The software is available for download at `http://www.stellman-greene.com/FindRelated/`. Prior research leveraging the intellectual linkages between articles generated by PMRA include Azoulay et al. (2015), Azoulay et al. (forthcoming), and Myers (2018).

of these cutoff rules. Reassuringly, our results are qualitatively similar regardless of the rule employed.

To fix ideas, consider "The transcriptional program of sporulation in budding yeast" [PubMed ID 9784122], an article published in the journal *Science* in 1998 originating from the laboratory of Ira Herskowitz, an eminent UCSF biologist who died in 2003 from pancreatic cancer. As can be seen in Appendix Figure C4, PMRA returns 72 original related journal articles for this source publication. Some of these intellectual neighbors will have appeared before the source to which they are related, whereas others will have only been published after the source. Some will represent the work of collaborators, past or present, of Herskowitz's, whereas others will represent the work of scientists in her field she may never have come in contact with during her life, much less collaborated with. The salient point is that nothing in the process through which these related articles are identified biases us towards (or away from) articles by collaborators, frequent citers of Herskowitz's work, or co-located researchers.

Consider now the second most-related article to Herskowitz's *Science* paper listed in Figure C4, "Phosphorylation and maximal activity of *Saccharomyces cerevisiae* meiosis-specific transcription factor Ndt80 is dependent on Ime2." Figure C5 in Appendix C displays the MeSH terms that tag this article along with its source. As a byproduct, PMRA also provides a cardinal dyadic measure of intellectual proximity between each related article and its associated source article. In this particular instance, the relatedness score of "Phosphorylation..." is 94%, whereas the relatedness score for the most distant related article in Figure C4, "Catalytic roles of yeast..." is only 62%.

In the five years prior to his death (1998-2002), Herskowitz was the last author on 12 publications, the publications most closely associated with his position as head of a laboratory. For each of these source publications, we treat the set of publications returned by PMRA as constituting a distinct subfield, and we create a subfield panel dataset by counting the number of related articles in each of these subfields in each year between 1975 and 2006. An important implication of this data construction procedure is that the subfields we delineate are quite limited in scope. One window into the degree of intellectual breadth for subfields is to gauge the overlap between the articles that constitute any pair of subfields associated with the same star. In the sample, the 452 deceased stars account for 3,076 subfields, and 21,661 pairwise combination of subfields (we are only considering pairs of subfields associated with the same individual star). Appendix Figure C6 displays the histogram for the distribution of overlap, which is extremely skewed. A full half of these pairs exhibit exactly zero overlap, whereas the mean of the distribution is 0.06. To find pairs of subfields that display substantial amounts of overlap (for example, half of the articles in subfield 1 also belong in subfield 2), one must reach far into the right tail of the distribution, specifically, above the 98^{th} percentile.

As such, the subfields we delineate are relatively self-contained. Performing the analysis at the level of the subfield—rather than lumping together all the subfields of an individual star—will provide us with an opportunity to exploit variation in the extent of participation of the star within each of her subfields. We will also check the validity of the main results when rolling the data up from the subfield level to the star level in Appendix F. Finally, since even modest amounts of overlap entail that the observations corresponding to the subfields of individual stars will not be independent in a statistical sense, we will cluster standard errors at the level of the star scientist.[9]

3.3 Identification Strategy

Given our interests in the effect of superstar death on entry into scientific subfields, our empirical strategy is focused on changes in published research output after the superstar passes away, relative to when she was still alive. To ensure that we are estimating the effect of interest and not some other influence that is correlated with the passage of time, our specifications include age and period effects, as is the norm in studies of scientific productivity (Levin and Stephan 1991). These temporal controls are tantamount to using subfields that lost a superstar in earlier or later periods as an implicit control group when estimating entry into subfields that currently experienced the death of a superstar. If the death of a superstar only represented a one-time shift in the level of entry into the relevant subfields, this would not be problematic. But if these unfortunate events affect trends—and not simply levels—of scientific activity, this approach may not suffice to filter out the effect of time-varying omitted variables, even when flexible age and calendar time controls are included in the econometric specification. One tangible concern about time-varying effects relates to the life cycle of subfields, where productive potential may initially increase over time before peaking and then slowly declining.

[9]The compactness of these subfields likely reflect the technology of research within the life sciences, a similar exercise performed in a different domain of science, particularly those characterized by large collaborative projects, might well result in subfields with substantially more overlap.

To mitigate this threat to identification, our preferred empirical strategy relies on the selection of a matched scientist/subfield for each treated scientist/subfield. These control observations are culled from the universe of subfields in which superstars who do not die are active (see Section 3.1 and Appendix D). Combining the treated and control samples enables us to estimate the effect of superstar death in a difference-in-differences framework. Appendix Figure D1 illustrates the procedure used to identify control subfields in the particular case of the Herskowitz publication highlighted above.

We begin by looking at all the articles that appeared in the same journal and in the same year as the treated source articles. From this set of articles, we keep only those that have one of the still-living superstars in the last authorship position. Then, using a “coarsened exact matching” procedure detailed in Appendix D, the control source articles are selected such that (1) the number of authors in the treated and control are approximately similar; (2) the age of the treated and control superstars differ by no more than five years; and (3) the number of citations received by the treated and source article are similar. For the Herskowitz/“sporulation in budding yeast” pair, we can select 10 control articles in this way. All of these controls were also published in *Science* in 1998, and have between five and seven authors. One of these controls is “Hepatitis C Viral Dynamics in Vivo...,” whose last author is Alan Perelson, a biophysicist at Los Alamos National Lab. Perelson and Herskowitz obtained their PhD only a year apart. The two papers had received 514 and 344 citations respectively by the end 2003. Though this is a large difference, this places both well above the 99^{th} percentile of the citation distribution for 5-year old articles published in 1998.

One potential concern with the addition of this “explicit” control group is that control subfields could be affected by the treatment of interest. What if, for instance, a control source article happens to be related (in a PMRA sense) with the treated source? Because the subfields identified by PMRA are narrow, this turns out to be very infrequent. Nonetheless, we remove all such instances from the data. We then find all the intellectual neighbors for these control source articles using PMRA; a control subfield is defined by the set of related articles returned by PMRA, in a manner that is exactly symmetric to the procedure used to delineate treated subfields. When these related articles are parsed below to distinguish between those published by collaborators and non-collaborators of the star, or between those by intellectual outsiders and insiders, covariates for treated and control observations will always be defined with perfect symmetry.

3.4 Descriptive Statistics

The procedure described above yields a total of 34,218 distinct subfields; 3,076 subfields correspond to one of the 452 dead scientists, whereas 31,142 subfields correspond to one of 5,809 still-living scientists. Table 2 provides descriptive statistics for control and treated subfields in the baseline year, i.e., the year of death for the deceased scientist.[10]

Covariate balance. In the list of variables displayed in Table 2, a number of covariates are balanced between treated and control subfields solely by virtue of the coarsened exact matching procedure—for instance, (star) investigator year of degree, the source article number of authors, or the source article number of citations at baseline. However, there is nothing mechanical to explain the balance between treated and control subsamples with respect to the stock of our main outcome variable: the number of articles in the star's field. Figure 1 compares the distributions of the cumulative number of articles published in our sample of subfields up to the year of death, broken down by treatment status. Overall, one can observe a great deal of overlap between the two histograms; the means and medians are virtually identical. Of course, balance in the levels of the outcome variable is not technically required for the validity of the empirical exercise.[11] Yet, given the *ad hoc* nature of the procedure used to identify control subfields, this degree of balance is reassuring.

Another happy byproduct of our matching procedure is that treated and control scientists also appear quite similar in the extent of their eminence at the time of (counterfactual) death, whether such eminence is measured through NIH funding, the number of articles published, or the number of citations these articles received.

Collaborators vs. non-collaborators. One critical aspect of the empirical analysis is to distinguish between collaborators and non-collaborators of the star when measuring publishing activity in a subfield. It is therefore crucial to describe how this distinction can be made in our data. Information about the superstars' colleagues stems from the Faculty Roster of the Association of American Medical Colleges (AAMC), to which we secured licensed access for the years 1975 through 2006, and which we augmented using NIH grantee information (cf. Azoulay et al. [2010] for more details).

[10] We can assign a counterfactual year of death for each control subfield, since each control subfield is associated with a particular treated subfield through the matching procedure described above.

[11] What is required is that the trends in publication activity be comparable between treated and control subfields up until the death of the treated scientist. We verify that this is the case below.

An important implication of our reliance on these sources of data is that we can only identify authors who are faculty members in U.S. medical schools, or recipients of NIH funding. We cannot systematically identify scientists working for industrial firms, or scientists employed in foreign academic institutions.[12] The great benefit of using AAMC data, however, is that they ensure we have at our disposal both demographic and employment information for every individual in the relevant labor market: their (career) age, type of degree awarded, place of employment, gender, and research output, whether measured by publications or NIH grants.

To identify authors, we match the authorship roster of each related article in one of our subfields with the AAMC roster.[13] We tag as a collaborator any author who appeared as a co-author of the star associated with the subfield on any publication prior to the death. Each related article is therefore assigned to one of two mutually-exclusive bins: the "collaborator" bin comprises the set of publications with at least one identified author who coauthored with the star prior to the year of death (or counterfactual death); the "non-collaborator" bin comprises the set of publications with no identified author who coauthored with the star prior to the year of death (or counterfactual death).[14] As can be seen in Table 2, roughly 11% of the publication activity at baseline can be accounted for by collaborators. Moreover, this proportion is very similar for control and treated subfields.[15]

A first look at subfield activity. Figure E1 in Appendix E confirms that the treated and control subfields are on similar trajectories in publication activity up to the time of superstar death (though they diverge after the death event). This provides suggestive evidence for the validity of our research design, and is notable since the coarsened exact matching procedure that generated the sample of control subfields did not make any use of these outcomes. Moreover, the absence of differential trends can be observed for overall activity, for activity restricted to collaborators of the star, and for the publishing activity of non-collaborators.

[12] We can identify trainees who later go on to secure a faculty position, but not those who do not stay in academia.

[13] We limit ourselves to authors with relatively infrequent names. Though this may create some measurement error, there is no reason to suspect that the wrongful attribution of articles to authors will impact treated and control subfields in a differential way.

[14] We identify the publications in the subfield for which the superstar is an author and eliminate them from these calculations. As a result, any decrease in activity within the subfield cannot be ascribed to the mechanical effect of its star passing away.

[15] We define collaboration status by looking at the authorship roster for the entire corpus of work published by the star before or in the year of death, and not only with respect to the articles of the star that belong to the focal subfield.

More boldly, we can use these averages in the raw data to examine changes in outcomes after the death. For both treated and control subfields, the curves exhibit a pronounced inverted U-shaped pattern, with activity first increasing until it reaches a peak roughly two years before the death of the star (or counterfactual death for the control subfields and their associated stars). Activity then decreases steadily, but the slope of the decrease appears more pronounced for control subfields, relative to treated subfields (Panel A). This pattern is flipped when examining activity due to collaborators (Panel B): the relative decline is much more pronounced for treated subfields, which is consistent with the results in Azoulay et al. (2010). Panel C, which focuses on subfield activity limited to non-collaborators, provides the first non-parametric evidence that the downward-sloping part of the activity curve is less steep for treated subfields.

Figure E1 provides a transparent illustration of subfield publication activity over time, which proceeds directly from averaging the raw data, but the evidence it provides should be handled with an abundance of caution. First, it conflates calendar time and experimental time, when in actuality the death events in the data occur at varying frequencies between the years 1975 and 2003. Second, covariates like field age are not perfectly balanced across the treated and control groups, since the number of control subfields is not identical across treated subfields. Finally, it abstracts away from robust inference, and particularly from clustering: one would expect the subfield outcomes associated with an identical star to be correlated. Our econometric framework, described below, addresses these limitations and as a result provides a more solid foundation for the estimation of the causal effect of star death on the dynamics of subfield activity.

4 Results

The exposition of the econometric results proceeds in stages. After a review of methodological issues, we provide results that pertain to the main effect of superstar death on subfield growth, measured by publication rates and funding flows. Next, we attempt to elucidate the mechanism (or set of mechanisms) at work to explain our most robust finding, that of relative subfield growth in the wake of a star's passing, a growth entirely accounted for by contributions from non-collaborators. We do so by examining the characteristics of the articles published by non-collaborators, before turning to the characteristics of their authors. We also explore heterogeneity in the treatment effect through the interaction of the post-death indicator variable with various attributes of the stars and the subfields.

4.1 Econometric Considerations

Our estimating equation relates publication or funding activity in subfield i in year t to the treatment effect of losing a superstar:

$$E\left[y_{it}|X_{it}\right] = exp\Big[\beta_0 + \beta_1 AFTER\_DEATH_{it} + \beta_2 AFTER\_DEATH_{it} \times TREAT_i + f(AGE_{it}) + \delta_t + \gamma_i\Big] \quad (1)$$

where y is a measure of subfield activity, $AFTER\_DEATH$ denotes an indicator variable that switches to one in the year after the superstar associated with i passes away, $TREAT$ is an indicator variable for treated subfields, $f(AGE_{it})$ corresponds to a flexible function of the field's age, the δ_t's stand for a full set of calendar year indicator variables, and the γ_i's correspond to subfield fixed effects, consistent with our approach to analyze *changes* in activity within subfield i following the passing of a superstar.[16]

The subfield fixed effects control for many time-invariant characteristics that could influence research activity, such as the need for capital equipment or the extent of disease burden (e.g., for clinical fields). A pregnant metaphor for the growth of scientific knowledge has been that of biological evolution (Hull 1988; Chavalarias and Cointet 2013): a field is born when new concepts are introduced, resulting in an accelerating production of "offspring" (articles), until the underlying scientific community loses its thematic coherence, ushering in an era of decline (or alternatively, splitting or merging events). To flexibly account for such life cycle effects, we include subfield age indicator variables (where subfield age is computed as the number of years since the year of publication for the source article). The calendar year effects filter out the effects of the general expansion of the scientific enterprise as measured by the number of journals and articles published each year.[17]

We follow Jaravel et al. (2018) in including in our specification an indicator for the timing of death that is common to treated and control subfields (whose effect will be identified by the coefficient β_1) in addition to the effect of interest, an interaction between $AFTER\_DEATH$ and $TREAT$ (whose effect will be identified by the coefficient β_2). The effects of these two variables are separately identified because (i) death events are staggered across our observation period and (ii) control subfields inherit a counterfactual date of death because they

[16] To avoid confusion, we have suppressed any subscript for the superstars. This is without loss of generality, since each subfield is uniquely associated with a single star.

[17] It is not possible to separately identify calendar year effects from age effects in the "within subfield" dimension of a panel in a completely flexible fashion, because one cannot observe two subfields at the same point in time that have the same age but were born in different years (Hall et al. 2007).

are uniquely associated with a treated subfield through the matching procedure described in section 3.3. The inclusion of the common term addresses the concern that age, calendar year, and subfield fixed effects may not fully account for shifts in subfield activity around the time of the star's passing. If this is the case, $AFTER\_DEATH$ will capture the corresponding transitory dynamics, while $AFTER\_DEATH \times TREAT$ will isolate the causal effect of interest. Empirically, we find that in some specifications, the common term has substantial explanatory power, though its inclusion does not radically alter the magnitude of the treatment effect.

Estimation. The dependent variables of interest, including publication counts and NIH grants awarded, are skewed and non-negative. For example, 31.40% of the subfield/year observations in the data correspond to years of no publication activity; the figure climbs to 56.70% if one focuses on the count of NIH grants awarded. Following a long-standing tradition in the study of scientific and technical change, we present conditional quasi-maximum likelihood (hereafter QML) estimates based on the conditional fixed effects Poisson model developed by Hausman et al. (1984). Because the Poisson model is in the linear exponential family, the coefficient estimates remain consistent as long as the mean of the dependent variable is correctly specified (Gouriéroux et al. 1984).

Inference. QML (i.e., "robust") standard errors are consistent even if the underlying data generating process is not Poisson. In fact the Hausman et al. estimator can be used for any non-negative dependent variables, whether integer or continuous (Santos Silva and Tenreyro 2006), as long as the variance/covariance matrix is computed using the outer product of the gradient vector (and therefore does not rely on the Poisson variance assumption). Further, QML standard errors are robust to arbitrary patterns of serial correlation (Wooldridge 1997), and hence immune to the issues highlighted by Bertrand et al. (2004) concerning inference in DD estimation. We cluster the standard errors around superstar scientists in the results presented below.[18]

[18] Knowledge spillovers and scientific breakthroughs, including the adoption of research tools, could encourage innovation across related fields. This possibility is not entirely dealt with by clustering inference at the star level, since spatial dependence in knowledge space could occur between any pair of subfields, whereas clustering only allows for dependence among the subfields associated with the same star. As it turns out, the Poisson conditional fixed effects estimator also provides a consistent estimator of the variance in the presence of time-invariant patterns of spatial auto-correlation (Bertanha and Moser 2016).

Dependent Variables. Our primary outcome variable is publication activity in a subfield. However, we go beyond this raw measure by assigning the related articles that together constitute the subfield into a variety of bins. For instance, we can decompose publication activity in the subfield into two mutually exclusive subfields: articles with a superstar on the authorship roster vs. articles without a superstar; etc. Articles in each bin can then be counted and aggregated up to the subfield/year level.

Capturing funding flows at the field level is slightly more involved. *PubMed* systematically records NIH grant acknowledgements using grant numbers. Unfortunately, these grant numbers are often truncated and omit the grant cycle information that could enable us to pin down unambiguously the particular year in which the grant was awarded. When it is missing, we impute the award year using the following rule: for each related publication that acknowledges NIH funding, we identify the latest year in the three-year window that precedes the publication during which funding was awarded through either a new award or a competitive renewal. To measure funding activity in a subfield, we create a count variable that sums all the awards received in particular year, where these awards ultimately generate publications in the focal subfield.

4.2 Main effect of superstar death

Table 3 and Figure 2 present our core results. Overall, we find that publication activity increases slightly following the death of a star scientist who was an active contributor to it, but the magnitude of the effect is modest (about 5.2%) and imprecisely estimated (column 1). Yet, this result conceals a striking pattern that is uncovered when we distinguish between publications by collaborators and non-collaborators. The decline in publication activity accounted for by previous collaborators of the star is large, on the order of 20.7% (column 2). This evidence is consistent with previous findings, which showed that coauthors of superstar scientists who die suffer a drop in output, particularly if their non-collaborative work exhibited strong keyword overlap with the star, i.e., if they were intellectually connected in addition to being coauthors (Azoulay et al. 2010, Table VI, column 2).

A limitation of the previous work focusing on the fate of collaborators after the loss of an eminent scientist always lied in the failure to distinguish between social and intellectual channels of influence, since every treated scientist was by definition a collaborator, even if merely a casual one. In this study, we can relax this constraint, and when we do, we find that relative publication activity by non-collaborators in the subfield increases by a statistically significant $100 \times (e^{0.082} - 1) = 8.6\%$ (column 3).[19]

We also explore the dynamics of the effects uncovered in Table 3. We do so by estimating a specification in which the treatment effect is interacted with a set of indicator variables corresponding to a particular year relative to the superstar's death, and then graphing the effects and the 95% confidence interval around them (Panels A, B, and C of Figure 2 correspond to columns 1, 2, and 3 in Table 3).[20]

Two features of the figure are worthy of note. First, the dynamics amplify the previous results in the sense that we see the effects increasing (in absolute value) monotonically over time—there is no indication that the effects we estimated in Table 3 are merely transitory. Five years after a star's death, the relative increase in publication activity by non-collaborators is large enough in magnitude to fully offset the decline in activity by collaborators. Second, there is no discernible evidence of an effect in the years leading up to the death, a finding that validates *ex post* our identification strategy.

Nevertheless, the case for the exogeneity of death events with respect to the course of knowledge growth and decline within a subfield is stronger for sudden causes of deaths than for anticipated causes of death. Figure E2 in Appendix E provides a version of Figure 2, Panel C (event study graphs for non-collaborators) broken down by causes of death (anticipated vs. sudden). While there is more variability in the estimated path of outcomes in the years leading up to the death event in the anticipated case (Panel A) than in the sudden case (Panel B), it is imprecisely estimated and non-monotonic. In both panels, however, one can observe a slow but steady increase after the event in the rate of contributions by non collaborators in treated subfields, relative to control subfields. The distinction between sudden and anticipated events is explored further in section 4.4.

The last three columns of Table 3 focus on funding flows from the National Institutes of Health (NIH) rather than publication flows. More precisely, the outcome variable in columns 4, 5, and 6 is the number of distinct NIH awards that acknowledge a publication in the subfield in the three-year window before the year of publication for the related article

[19] The number of observations varies ever so slightly across columns because the conditional fixed effects specification drops observations corresponding to subfields for which there is no variation in activity over the entire observation period. This is true as well for the results reported in Tables 4 through 8.

[20] In these specifications, the $AFTER\_DEATH$ term which is common to treated and control subfields is also interacted with a complete series of lags and leads relative to the year of death or counterfactual death.

(summing the financial total of grant amounts, as opposed to the number of grants, yields similar results). The patterns are very similar to those obtained in the case of publication activity, both in terms of magnitudes and in terms of statistical significance.

4.3 Subfield growth patterns

In the remainder of the manuscript, we seek to characterize the kind of contribution, and the type of investigators that give rise to the novel empirical regularity we uncovered: that of relative growth for subfields following the death of their superstar anchor, a phenomenon entirely accounted for by research activity undertaken by scientists who never collaborated with the star while alive. As a consequence, all the results below pertain to contributions by non-collaborators; any article with even one author who collaborated with the star is excluded from the count of articles that constitute the dependent variable.

The impact and direction of new research. What characterizes the additional contributions that together lead to increased activity in a subfield after a star has passed on? Are these in fact important contributions to the subfield? Do they continue to focus on mainstream topics within the subfield, or should they be understood as taking the intellectual domain in a novel direction? Tables 4 and 5 explore these issues.

In Table 4, we parse every related article in the subfields to assign them into one of six mutually exclusive bins, based on their vintage-specific long-run citation impact: articles that fall in the bottom quartile of the citation distribution; in the second quartile; in the third quartile; articles that fall above the 75^{th} percentile, but below the 95^{th} percentile; articles that fall above the 95^{th} percentile, but below the 99^{th} percentile; articles that fall above the 99^{th} percentile of the citation distribution.[21] Each column in Table 4 (with the exception of the first which simply replicates the effect for all papers, regardless of impact, that was previously displayed in Table 3, column 3) reports the corresponding estimates. A startling result is that the magnitude of the treatment effect increases sharply and monotonically as

[21] A vintage is comprised of all the articles published in a given year. When we are referring to the vintage-specific, article-level distribution of citations, the relevant universe to compute quantiles is not limited to the articles that constitute the subfields in our data. Rather, the relevant universe includes the entire set of 17,312,059 articles that can be cross-linked between *PubMed* and the Web of Science. As a result, there is no reason to suspect that individual stars, or even our entire set of stars, could ever alter the shape of these distributions. For example, the article by Sopko et al. highlighted on Figure C5 (in Appendix C) received 40 citations from other articles in *PubMed* by 2015. This puts this article above the 79^{th} percentile of the citation distribution for articles published in 2002.

we focus on the rate of contributions with higher impact. In contrast, the number of lower-impact articles contributed by non-collaborators contracts slightly, though the effect is not precisely estimated.[22]

Table 5 parses the related articles in each subfield to ascertain whether contributions by non-collaborators constitute a genuine change in intellectual direction. Panel A distinguishes between contributions that are proximate in intellectual space to the source article from those that are more distant (though still part of the subfield as construed by PMRA). Because we have at our disposal both a cardinal and an ordinal measure of intellectual proximity, we present two sets of estimates. In both cases, the magnitude of the treatment effect pertaining to PMRA-proximate publication activity is larger, and more precisely estimated than the magnitude corresponding to PMRA-distant publication activity (relative to the same patterns for the control group of subfields). We can certainly rule out the conjecture that non-collaborators enter the field from the periphery. Rather, their contributions appear to tackle mainstream topics within the subfield.

Panel B sheds light on the intellectual direction of the field, by examining the cited references contained in each related article. The first two columns separate related articles in two groups: publications that cite at least some work which belongs to the subfield identified by PMRA for the corresponding source and publications that cite exclusively out of the PMRA subfield. Only articles in the second group appear to experience growth in the post-death era. The next two columns proceed similarly, except that the list of references is now parsed to highlight the presence of articles authored by the star (Column 3), as opposed to all other authors (Column 4). We find that subfield growth can be mostly accounted for by articles from non-collaborators who do not build on the work of the star.

Whereas Panel B highlighted the extent to which contributors were bringing new sources of inspiration into the subfield, Panel C focuses on the extent to which the treated subfields move closer to the scientific frontier in the wake of the superstar's passing. The first two columns do so by distinguishing between contributions that draw on recent versus more

[22] Table E3 and Figure E3 in Appendix E break down these results further by examining separately the growth of subfields by cause of death (anticipated vs. sudden). As mentioned earlier, the case for exogeneity is stronger for sudden death, since when the death is anticipated, it would be theoretically possible for the star to engage in "intellectual estate planning," whereby particular scientists (presumably close collaborators) are anointed as representing the next generation of leaders in the subfield. Our core results continue to hold when analyzed separately by cause of death. However, we gain statistical power from pooling these observations, and some empirical patterns would be estimated less precisely if we chose to focus solely on observations corresponding to subfields for which the star died suddenly and unexpectedly.

dated references. This exercise is repeated in Columns 3 and 4, with a focus on the vintage of the MeSH term combinations for each article in the subfield.[23] Both sets of results indicate that these new contributions are more likely to build on science of a more recent vintage.

Taken together, the results presented in Table 5 paint a nuanced picture of directional change in the wake of superstar passing. The new contributions do not represent a radical departure from the subfield's traditional concerns (Panel A). At the same time, the citation and MeSH evidence (Panels B and C) make it clear that these additional contributions are more likely to draw on new-to-the-subfield as well as new-to-the-world ideas. In short, they both rejuvenate the subfield, and alter its angular velocity by shifting its intellectual center of gravity away from its pre-death position.

It is important to note, however, that the findings above do not imply that the published results of entrants necessarily contradict or overturn the prevailing scientific understanding and assumptions within a subfield. We provide indirect evidence regarding these contributions' disruptive impact by leveraging a measure recently proposed by Funk and Owen-Smith (2017). Their index captures the degree to which an idea consolidates or destabilizes the status quo, by measuring whether the future ideas that build on the focal idea also rely on its acknowledged predecessors. The results in Table E4 of Appendix E suggest that these contributions do not radically disrupt the subfield. Rather, they appear to reflect the impact of a myriad "small r," permanent revolutions whereby new ideas come to the fore without necessarily eclipsing prior approaches.

Outsiders vs. competitors. The next step of the analysis is to investigate the type of scientists who publish the articles that account for subfield growth in the wake of a star's death. We examine the proximity in intellectual space between non-collaborators in the subfield and the deceased superstar. One possibility is that non-collaborators are competitors of the star, with much of their publication activity falling into the subfield when the star was alive. Another possibility is that they are recent entrants into the subfield—intellectual outsiders. To distinguish these different types of authors empirically, we create a metric of intellectual proximity for each related author we can match to the AAMC Faculty Roster, by computing the fraction of their publications that belongs to the star's subfields up to

[23] A two-way MeSH term combination is born in the year where an article is annotated by the keyword pair for the first time.

the publication year for each related article.[24] The distribution of this field overlap measure is displayed on Panel A of Figure 3. The distribution is skewed, with a pronounced mass point at the origin: approximately 50% of the related articles turn out to have authors with exactly zero intellectual overlap with the star's subfield, and another 1.24% are authored by new scientists for whom this publication within the subfield is also their first publication overall.

We now use this metric to gauge the extent to which the post-death publication activity by non-collaborators (relative to the control group) can be attributed to related authors whose outsider status falls into one of twelve separate bins. This includes one bin for new scientists, one bin for the bottom half of the overlap distribution, one bin for every five percentiles above the median (50^{th} to 55^{th} percentile, 55^{th} to 60^{th} percentile,..., 95^{th} to 99^{th} percentile), as well as a top percentile bin. We then compute the corresponding measures of subfield activity by aggregating the data up to the subfield/year level. These results are presented graphically in Panel B of Figure 3. Each dot corresponds to the magnitude of the treatment effect in a separate regression with the outcome variable being the number of articles in each subfield that belong to the corresponding bin.

A striking pattern emerges. The authors driving the growth in relative publication activity following a star's death are largely outsiders. They do not appear to have been substantially active in the subfield when the star was alive. In other words, they are predominantly new entrants into these subfields, though not necessarily novice scientists.

4.4 The Nature of Entry Barriers

The evidence so far points to fields of deceased stars enjoying bursts of activity after the death event. The influx of outsiders documented above suggests that stars may be able to regulate entry into their field while alive. In this section, we attempt to uncover the precise nature of barriers to entry into the subfields where the stars were prominent prior to their untimely demise. Methodologically, we do so by splitting the sample of fields across the median for a series of relevant covariates. Because there is no presumption that death events are exogenous with respect to subfield growth and decline within the strata

[24] Whenever we match more than one author on a related article, we assign to that article the highest proximity score for any of the matched authors. Appendix E, Table E9 defines overlap with respect to all the subfields associated with a given star, rather than simply the focal subfield. This does not alter our conclusions.

delineated by these covariates, it should be clear that we will only be able to document conditional correlations, and not causal effects in what follows.[25]

While it is tempting to envisage conscious effort by the stars to block entry through the explicit control of key resources, such as funding and/or editorial goodwill (Brogaard et al. 2014; Li 2017), this explanation appears inconsistent with the facts on the ground. In the five-year window before death, only three of our stars (out of 452) were sitting on study sections, the funding panels that evaluate the scientific merits of NIH grant applications. Another three were journal editors in the same time window. This handful of individuals could not possibly drive the robust effects we have uncovered.[26] If barriers to entry are not the result of explicit control by stars, what is discouraging entry?

Goliath's shadow. One possibility is that outsiders are simply deterred by the prospect of challenging a luminary in the field. The existence of a towering figure may skew the cost-benefit calculations from entry by outside scholars toward delay or alternative activities. Table 6 examines this role of implicit barriers to entry by focusing on the eminence of the star. Eminence is measured through the stars publication count, the stars cumulative number of citations garnered up to the year of death, and the stars cumulative amount of NIH funding. We also have a "local" measure of eminence: the star's importance to the field, which is defined as the fraction of papers in the subfield that have the star as an author. Splitting the sample at the median of these measures reveals a consistent pattern of results. Stars that were especially accomplished appear to be an important deterrent to entry, with their passing creating a larger void for non-collaborators to fill. Rather than directly thwarting the efforts of potential entrants, it appears that the mere presence of a preeminent scholar is sufficient to dissuade intellectual outsiders from engaging with the field.

Of course, the accomplishment of the star alone may not be the only factor influencing entry. We next turn our attention to how the characteristics of the field and the star's coauthors may also modulate this relationship. Since entry is largely confined to those fields that have lost an eminent star, the analysis that follows limits attention to those subfields

[25] Instead of interacting the treatment effect with covariates, we prefer to estimate our benchmark specifications on subsamples corresponding to below and above the median of these covariates. For these two approaches to yield comparable results, one would need to also saturate the specification with interaction terms between the covariates and year/field age effects. In practice, we have found that the fixed effects Poisson models fail to converge with this full set of interactions.

[26] We verified that omitting these scientists from the sample hardly change the core results.

in which the most eminent among the stars were active, as measured by our citation metric in Table 6.[27]

Subfield coherence. Entry into a field, even after it has lost its star, may be deterred if the subfield appears unusually coherent to outsiders. A subfield is likely to be perceived as *intellectually* coherent, when the researchers active in it agree on the set of questions, approaches, and methodologies that propel the field forward. Alternatively, a field might be perceived as *socially* coherent, when the researchers active in it form a tightly-knit clique, often collaborating with each other, and perhaps also reviewing each other's manuscripts. To explore these purported barriers to subfield entry, we develop two alternative measures of intellectual coherence, and one measure of social coherence.

Our first index of intellectual coherence leverages PMRA to capture the extent to which articles in the subfield pack themselves into a crowded scientific neighborhood. Recall that for each article in a subfield, we have at our disposal both a cardinal and an ordinal measure of intellectual proximity with the source article from which all other articles in the subfield radiate. Focusing only on the set of articles published in the subfield before the year of death, we measure intellectual coherence as the cardinal ranking (expressed as a real number between zero and one) for the 25^{th} most related article in the subfield.[28] According to this metric, subfields exhibit wide variation in their degree of intellectual coherence, with a mean and median equal to 0.60 ($sd = 0.13$). The second index of intellectual coherence exploits the list of references cited in each article in the subfield before the star's death. In the spirit of Funk and Owen-Smith (2017), for all related articles published in the five years prior to the star's death, we compute the fraction of references that fall within the subfield. Our contention is that subfields that are more self-referential will tend to dissuade outsiders from entering. Once again, we observe meaningful variation across subfields using this second index ($mean = 0.05; sd = 0.04$).

[27]More precisely, Table 7 below drops from the sample subfields associated with stars who fall below the median of cumulative citations garnered by the year of death. Results are qualitatively similar when focusing on the most eminent stars as defined by publications or NIH funding. Table F6 in Appendix F presents the results corresponding to the subsample of less-eminent stars.

[28]The choice of the twenty fifth-ranked article is arbitrary, and also convenient. After purging from each subfield reviews, editorials, and articles appearing in journals not indexed by *WoS*, 95% of the subfields contain 25 articles or more in the period that precedes the star's death. In those rare cases where the number of articles is less than twenty-five, we choose as our measure of coherence the cardinal measure for the least-proximate article in the subfield.

Our measure of social coherence summarizes the degree of "cliquishness" within a subfield by computing the clustering coefficient in its coauthorship network. The clustering coefficient is simply the proportion of closed triplets within the network, an intuitive way to measure the propensity of scientists in the field to choose insiders as collaborators.[29]

Panel A of Table 7 investigates the role of these intellectual and social barriers in modulating the post-death expansion of fields. We find tentative evidence of a role for both types of barriers, in that the magnitude of the treatment effect for coherent fields is always smaller than the magnitude for less coherent fields, regardless of how coherence is measured. The *difference* between the estimates for more or less coherent subfields does not reach statistical significance at conventional levels. What seems notable, however, is that the magnitudes are consistently ordered across the three measures.

Incumbent resource control. While we noted earlier that stars do not appear especially well positioned to directly block entry through the control of key resources, it is possible that those resources can be controlled indirectly through the influence of collaborators. If incumbent scholars within a field serve as gatekeepers of funding and journal access, they may be able to effectively stave off threats of entry from outsiders. The same may be implicitly true if collaborators are the recipients of the lion's share of funding within the field. To assess financial gatekeeping, we use information regarding the composition of NIH funding panels, to tabulate, for each star, the number of collaborators who were members of at least one of these committees in the five years preceding the death of the star. We would like to proceed in a similar fashion using the composition of editorial boards, but these data are not easily available for the set of *PubMed*-indexed journals and the thirty-year time period covered by our sample. As an alternative, we develop a proxy for editorial position based on the number of editorials or comments written by every collaborator of the star.[30] We then sum the number of editorials written by coauthors in the five years before the death. Together, the editorial and study section information allow us to distinguish between the

[29]The clustering coefficient is based on triplets of nodes (authors). A triplet consists of three authors that are connected by either two (open triplet) or three (closed triplet) undirected ties. The clustering coefficient is the number of closed triplets over the total number of triplets (both open and closed, cf. Luce and Perry [1949]).

[30]We investigated the validity of this proxy as follows. In the sample of deceased superstars, every individual with five editorials or more was an editor. In a random sample of 50 superstars with no editorials published, only one was an editor (for a field journal). Finally, among the sixteen superstars who wrote between one and four editorials over their career, we found two whose CV indicate they were in fact editors for a key journal in their field. We conclude that there appears to be a meaningful correlation between the number of editorials written and the propensity to be an editor.

stars whose coauthors were in a position to channel resources towards preferred individuals or intellectual approaches from those stars whose important coauthors had no such power.

Panel B of Table 7 presents the evidence on the role of indirect control. The results paint a consistent, if not always statistically significant, picture. While subfield expansion is the rule, it appears more pronounced when stars have relatively few collaborators in influential positions, or collectively capture a smaller portion of the funding that supported research in the subfield. Indirect control therefore appears to be a potential mechanism through which superstars can exert influence on the evolution of their fields, even from beyond the grave. Coauthors, either through their direct effort to keep the star's intellectual flame alive or simply by their sheer (financial) dominance in the field, erect barriers to entry into those fields that prevent its rejuvenation by outsiders.

Taken together, these results suggest that outsiders are reluctant to challenge hegemonic leadership within a field when the star is alive. They also highlight a number of factors that constrain entry even after she is gone. Intellectual, social, and resource barriers all impede entry, with outsiders only entering subfields whose topology offers a less hostile landscape for the support and acceptance of "foreign" ideas.

4.5 Welfare considerations

What are the implications of our results for welfare? We approach this question with a great deal of caution, since much of the evidence presented thus far pertains to changes in the *direction*, rather than the *rate*, of scientific progress. Making welfare statements in this context is tantamount to valuing the importance of the new directions in which related authors take their fields (compared to the prior agenda inherited from the superstar), as well as ascertaining the fate of fields that the new entrants departed, and the agenda they otherwise might have pursued had the star remained alive. Such an exercise is fraught with peril. Below we synthesize the results that already speak to these questions, and provide a number of additional pieces of evidence. Together, this collage of results builds a circumstantial case for the view that once securely ensconced at the helm of their field, stars leverage their power for longer than a benevolent social planner might prefer.

Our earlier evidence suggests that entrants bring different and more recent ideas into the subfields they enter to create highly impactful output (Tables 4 and 5). In Appendix E we further show that the subfields that experience the largest post-death boost in activity are those in which the star was presiding over an empire that was losing momentum in the years immediately preceding the star's death (Tables E5 and E8). These subfields are also those in which the star's close collaborators were less able to regulate entry (Table 7B).

It is important to note, however, that the additional output by entrants in treated subfields is largely offset by commensurate declines in output by the star's collaborators (Table 3). Moreover, these new contributions appear to come at the expense of the entrants' prior agenda. In Appendix G, we examine changes in total output at the related author level, using a difference-in-differences set-up that parallels our analyses at the subfield level. The results in Table G1 show that non-collaborators do not increase their overall output, measured in terms of publications and NIH grants awarded. Since we know from our main analysis that related authors are contributing more within the subfields of dead superstars, the absence of changes in total output imply that this additional work is displacing work they were doing in other subfields. Their new output replaces, at least in part, articles that these authors would have written in other intellectual domains had the star remained alive.[31]

As a whole, these results imply that entrants are moving subfields in productive directions relative to the period immediately preceding the passing of the star, but without increasing scientific output in the aggregate.

However, the impacts in the final years of a star's life are not necessarily indicative of their contributions writ large. Indeed, the lofty accomplishments which earned them superstar status suggest that their net contribution to society is likely positive. A longer view would also recognize that the scientific journeymen of today may well become the stars of tomorrow with a career that slowly builds to an apex of socially valuable accomplishments, that will someday experience a similar decline (see Figure E4 in Appendix E).

One lens into this phenomenon is to examine the status of scientists that produce new contributions in a subfield. In the first two columns of Table 8, we parse every article by non-collaborators, distinguishing between those that have a star author from those for which none of the authors are stars. We find that the effect is driven by related articles where none

[31]We also estimate a dynamic version of these specifications and display the corresponding event study-style graphs in Figure G1 (publication output) and Figure G2 (grant output). In general, it appears from these figures that the total output of related authors neither expands nor contracts in the wake of a star's passing.

of the authors is particularly famous. One limitation of this dichotomy is that it fails to take into account long-run career trajectories, since it lumps together mediocre scientists with those that have not yet made their mark, but will do so in the future.

We can explore this dynamic by taking advantage of the fact that roughly 20% of the eminent life scientists in our sample have a clear date attached to their accession to star status: the year of appointment as a Howard Hughes Medical Investigator, or the year of election to the National Academy of Science or the National Academy of Medicine. These events mark their recipients as among the most celebrated within the superstar sample. With this more rarefied definition of stardom, we can now distinguish between related authors who are "never stars," "current stars," and "future stars." The next three columns of Table 8 show that future star authors are disproportionately likely to contribute to treated subfields after the star has passed away, consistent with the idea that the outsiders of today can sometimes turn into the stars of tomorrow—a phenomenon we refer to as the circle of academic life.

In light of these and our earlier results, we refrain from drawing any strong welfare conclusions. An aggregate assessment of the value of stars would require us to integrate accomplishments over the lifecourse of stars and everyone else who followed in their footsteps, with a particular focus on the fate of the fields from which new entrants divest. That does not leave us completely empty handed. In the next section, we offer some modest policy recommendations that focus on accelerating idea churn in fields dominated by scientists in the twilight of their careers.

4.6 Extensions and robustness

Appendix E presents results pertaining to extensions of the main analyses. Appendix F provides a number of robustness checks. In the interest of space, we only call out a subset of the analyses presented therein, but we have written these appendices as stand-alone documents, such that the interested reader can consult them for additional details.

Impact of research infrastructure needs. Our analysis is limited to the life sciences. Though this area accounts for a large fraction of publicly funded, civilian research funding in the United States, it is not necessarily representative of all fields of science. In particular, some domains of research require access to expensive and specialized capital equipment. When capital needs are large and lumpy, the evolution of subfields in the wake of an eminent scientist's death will likely depend on the institutions that govern access to the scarce capital equipment.

Within biomedical research, large-scale clinical trials most closely—albeit imperfectly—resemble the characteristics of capital-intensive scientific fields. These require a large infrastructure of data collection, monitoring, and management, which is why these activities are often consolidated in large cooperative groups such as the AIDS Clinical Trials Group, the Children's Oncology Group, or the Framingham Heart Study. *PubMed* has a "publication type" field which allows us to identify the subfields that are clinical-trial intensive (10% of the subfields) versus those that are not (the remaining 90%). Table E6 replicates the results of Table 3 separately for these two subsamples. Although our ability to estimate statistically significant effects is limited by sample size, the magnitudes are very similar.

Impact of star age and experience. As explained earlier, we do not impose a strict age cutoff for the deceased star, we merely insist that they exhibit tangible signs of research activity, such as publishing original articles, obtaining NIH grants, and training students. Among our 452 departed superstars, the median age at death is 61, the seventy-fifth percentile 67, and the top decile 73. How do the core results change when the scientists who passed away at an advanced age are excluded from the sample? As can be observed in Table E7, the subfields of stars who passed away more prematurely are responsible for most of the effect. The effect for the fields associated with older stars is small in magnitude and imprecisely estimated. We chose to keep these older stars in the sample because a larger sample affords us opportunities to explore mechanisms without losing power to detect nuanced effects statistically.

Star level analyses. In Table F1, we probe the robustness of the core results presented in Table 3 after rolling up the data to the level of the star scientist (deceased or control). Recall that the treatment variable exhibits variation at the level of the star scientist, and not at the level of a single subfield. In this robustness check, we lump all related articles for each star together as if they belonged to a single subfield. The results in Table F1 are quite similar to those in Table 3, both in terms of magnitude and statistical significance. One exception is the coefficient on the effect of entry by collaborators, which is negative as expected, but smaller in magnitude, relative to the corresponding coefficient in Table 3. The corresponding event-study graphs, displayed in Figure F3, also display patterns fully consistent with those observed for our benchmark set of results. As explained in Section 3.2, we strongly prefer performing the analyses at the the subfield level, for two reasons. First, the subfields delineated by PMRA exhibit limited overlap (see Figure C6 in Appendix C), and as a result the within-star, between subfield variation in publication activity can be exploited meaningfully. Second, we can track the differential position of the star across the subfields in which she was active. The covariates that leverage these differences help us shed light on mechanisms, as in Tables 7, E5, and E8.

Alternate functional forms. In Table F2, we examine the sensitivity of our benchmark set of results to the choice of alternative functional forms. In the three columns to the left, we simply use the "raw" number of articles in the subfield as the outcome, and perform estimation by OLS. Of course, the estimates are not directly interpretable in terms of elasticities. At the mean of the data, however, the treatment effect in the third column implies that subfield entry by non-collaborating authors expands by $0.409/3.335 = 12.26\%$, which is not all that different from the 8.2% reported in Table 3. In the three columns to the right, we report results corresponding to OLS estimation, but this time with the outcome variables transformed using the inverse hyperbolic sine function (Burbidge et al. 1988). In this case, coefficient estimates can be interpreted as elasticities, as an approximation. They are quite similar once again to those reported in Table 3, except for the effect on entry by collaborators, which is smaller in magnitude.

5 Conclusion

In this paper, we leverage the applied economist's toolkit, together with a novel approach to delineate the boundaries of scientific fields, to explore the effect that the passing of an eminent life scientist exerts on the dynamics of growth—or decline—for the fields in which she was active while alive. We find that publications and grants by scientists that never collaborated with the star surge within the subfield, absent the star. Interestingly, this surge is not driven by a reshuffling of leadership within the field, but rather by new entrants that are drawn from outside of it. Our rich data on individual researchers and the nature of their scholarship allows us provide a deeper understanding of this dynamic.

In particular, this increase in contributions by outsiders appears to tackle the mainstream questions within the field but by leveraging newer ideas that arise in other domains. This intellectual arbitrage is quite successful—the new articles represent substantial contributions, at least as measured by long-run citation impact. Together, these results paint a picture of scientific fields as scholarly guilds to which elite scientists can regulate access, providing them with outsized opportunities to shape the direction of scientific advance in that space.

We also provide evidence regarding the mechanisms that enable the regulation of entry. While stars are alive, entry appears to be effectively deterred where the shadow they cast over the fields in which they were active looms particularly large. After their passing, we find evidence for influence from beyond the grave, exercised through a tightly-knit "invisible college" of collaborators (de Solla Price and Beaver 1966; Crane 1972). The loss of an elite scientist central to the field appears to signal to those on the outside that the cost/benefit calculations on the avant-garde ideas they might bring to the table has changed, thus encouraging them to engage. But this occurs only when the topology of the field offers a less hostile landscape for the support and acceptance of "foreign" ideas, for instance when the star's network of close collaborators is insufficiently robust to stave off threats from intellectual outsiders.

In the end, our results lend credence to Planck's infamous quip that provides the title for this manuscript. Yet its implications for social welfare are ambiguous. While we can document that eminent scientists restrict the entry of new ideas and scholars into a field, gatekeeping activities could have beneficial properties when the field is in its inception; it might allow cumulative progress through shared assumptions and methodologies, and the ability to control the intellectual evolution of a scientific domain might, in itself, be a prize that spurs much *ex ante* risk taking. Because our empirical exercise cannot shed light on these countervailing tendencies, we must remain guarded in drawing policy conclusions from our results. Yet, the fact that the presence of a tutelar figurehead can freeze patterns of participation into a scientific field increases the appeal of policies that bolster access to less established or less well-connected investigators. Examples of such policies include caps on the amount of funding a single laboratory is eligible to receive, "bonus points" for first-time investigators in funding programs, emeritus awards to induce senior scientists to wind down their laboratory activities, and double-blind refereeing policies (Kaiser 2011, Berg 2012, Deng 2015).

All of the evidence we have presented pertains to the academic life sciences. It is unclear how the lessons from that setting might apply to other fields inside the academy. In particular, when frontier research requires access to expensive and highly-specialized capital equipment—as is sometimes the case in the physical sciences—the rules governing access to that capital are likely to favor succession by insiders. At the other end of the spectrum, more atomistic fields where scientists generally work alone or in very small groups may evolve in a more frictionless manner. Whether our findings apply to industrial research and development is also an open question. In that setting, the choice of problem-solving approaches is guided by market signals (however imperfectly, cf. Acemoglu [2012]), and thus likely to differ from those selected under the more nuanced system of pecuniary and non-pecuniary incentives that characterizes academic research (Feynman 1999; Aghion, Dewatripont, and Stein 2008). Assessing the degree to which our results extend to other settings, and the reasons they might differ, represents a fruitful area for future research.

References

Aad, Georges et al. 2015. "Combined Measurement of the Higgs Boson Mass in *pp* Collisions at $\sqrt{}$s=7 and 8 TeV with the ATLAS and CMS Experiments." *Physical Review Letters* **114**(191803): 1-33.

Acemoglu, Daron. 2012. "Diversity and Technological Progress." In Josh Lerner, and Scott Stern (Eds.), *The Rate & Direction of Inventive Activity Revisited*, pp. 319-356. Chicago, IL: University of Chicago Press.

Aghion, Philippe, Mathias Dewatripont, and Jeremy C. Stein. 2008. "Academic Freedom, Private Sector Focus, and the Process of Innovation." *RAND Journal of Economics* **39**(3): 617-635.

Aghion, Philippe, and Peter Howitt. 1992. "A Model of Growth through Creative Destruction." *Econometrica* **60**(2): 323-351.

Akerlof, George, and Pascal Michaillat. 2017. "Beetles: Biased Promotion and Persistence of False Belief." NBER Working Paper #23523.

Azoulay, Pierre, Joshua Graff Zivin, and Jialan Wang. 2010. "Superstar Extinction." *Quarterly Journal of Economics* **125**(2): 549-589.

Azoulay, Pierre, Jeffrey L. Furman, Joshua L. Krieger, and Fiona Murray. 2015. "Retractions." *Review of Economics and Statistics*, **97**(5): 1118-1136.

Azoulay, Pierre, Danielle Li, Joshua S. Graff Zivin, and Bhaven N. Sampat. 2015. "Public R&D Investment and Private Sector Patenting: Evidence from NIH Funding Rules." Forthcoming, *Review of Economic Studies.* Also NBER Working Paper #20889.

Azoulay, Pierre, Danielle Li, and Bhaven N. Sampat. 2017. "The Applied Value of Public Investments in Biomedical Research." *Science* **356**(6333): 78-81.

Bertanha, Marinho, and Petra Moser. 2016. "Spatial Errors in Count Data Regressions." *Journal of Econometric Methods* **5**(1): 49-69.

Berg, Jeremy M. 2012. "Science Policy: Well-funded Investigators Should Receive Extra Scrutiny." *Nature* **489**(7415): 203.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. "How Much Should We Trust Differences-in-Differences Estimates?" *Quarterly Journal of Economics* **119**(1): 249-275.

Borjas, George J., and Kirk B. Doran. 2015a. "Which Peers Matter? The Relative Impacts of Collaborators, Colleagues, and Competitors." *Review of Economics and Statistics* **97**(5): 1104-1117.

Borjas, George J., and Kirk B. Doran. 2015b. "Cognitive Mobility: Labor Market Responses to Supply Shocks in the Space of Ideas." *Journal of Labor Economics* **33**(S1): S109-S145.

Börner, Katy, Chaomei Chen, and Kevin W. Boyack. 2003. "Visualizing Knowledge Domains." *Annual Review of Information Science and Technology* **37**(1): 179-255.

Bourdieu, Pierre. 1975. "La Spécificité du Champ Scientifique et les Conditions Sociales du Progrès de la Raison." *Sociologie et Sociétés* **7**(1): 91-118.

Bramoullé, Yann, and Gilles Saint-Paul. 2010. "Research Cycles." *Journal of Economic Theory* **145**(5): 1890-1920.

Brogaard, Jonathan, Joseph Engelberg, and Christopher Parsons. 2014. "Network Position and Productivity: Evidence from Journal Editor Rotations." *Journal of Financial Economics* **111**(1): 251-270.

Burbidge, John B., Lonnie Magee and A. Leslie Robb, 1988. "Alternative Transformations to Handle Extreme Values of the Dependent Variable." *Journal of the American Statistical Association* **83**(401): 123-127.

Chavalarias, David, and Jean-Philippe Cointet. 2013. "Phylomemetic Patterns in Science Evolution—The Rise and Fall of Scientific Fields." *PLoS one* **8**(2): e54847.

Cole, Jonathan R., and Stephen Cole. 1972. "The Ortega Hypothesis." *Science* **178**(4059): 368-375.

Crane, Diana. 1972. *Invisible Colleges: Diffusion of Knowledge in Scientific Communities.* Chicago, IL: University of Chicago Press.

Crick, Francis. 1970. "Central Dogma of Molecular Biology." *Nature* **227**(5258): 561.

de Solla Price, Derek J. 1963. *Little Science, Big Science*. New York: Columbia University Press.

de Solla Price, Derek J., and Donald D. Beaver. 1966. "Collaboration in an Invisible College." *American Psychologist* **21**(11): 1011-1018.

Deng, Boar. 2015. "NIH Ponders Emeritus Grants." *Nature* **518**(7538): 146-147.

Feynman, Richard P. 1999. *The Pleasure of Finding Things Out*. New York: Basic Books.

Fortunato, Santo and Darko Hric. 2016. "Community Detection in Networks: A User Guide." *Physics Reports* **659**: 1-44.

Foster, Jacob G., Andrey Rzhetsky, and James A. Evans. 2015. "Tradition and Innovation in Scientists' Research Strategies." *American Sociological Review* **80**(5): 875-908.

Funk, Russell J., and Jason Owen-Smith. 2017. "A Dynamic Network Measure of Technological Change." *Management Science* **63**(3): 791-817.

Gorham, Geoffrey. 1991. "Planck's Principle and Jeans's Conversion." *Studies in History & Philosophy of Science* **22**(3): 471-497.

Gouriéroux, Christian, Alain Montfort, and Alain Trognon. 1984. "Pseudo Maximum Likelihood Methods: Applications to Poisson Models." *Econometrica* **53**(3): 701-720.

Hall, Bronwyn H., Jacques Mairesse, and Laure Turner. 2007. "Identifying Age, Cohort and Period Effects in Scientific Research Productivity: Discussion and Illustration Using Simulated and Actual Data on French Physicists." *Economics of Innovation and New Technology* **16**(2): 159-177.

Hausman, Jerry, Bronwyn H. Hall, and Zvi Griliches. 1984. "Econometric Models for Count Data with an Application to the Patents-R&D Relationship." *Econometrica* **52**(4): 909-938.

Hull, David L. 1988. *Science as a Process.* Chicago, IL: University of Chicago Press.

Hull, David L., Peter D. Tessner, and Arthur M. Diamond. 1978. "Planck's Principle." *Science* **202**(4369): 717-723.

Jaravel, Xavier, Neviana Petkova, and Alex Bell. 2018. "Team-Specific Capital and Innovation." *American Economic Review* **108**(4-5): 1034-1073.

Jones, Benjamin F. 2009. "The Burden of Knowledge and the 'Death of the Renaissance Man': Is Innovation Getting Harder?" *Review of Economic Studies* **76**(1): 283-317.

Kaiser, Jocelyn. "Darwinism vs. Social Engineering at NIH." *Science* **334**(6057): 753-754.

Kuhn, Thomas S. 1970. *The Structure of Scientific Revolutions*. Chicago, IL: University of Chicago Press.

Levin, Sharon G., and Paula E. Stephan. 1991. "Research Productivity over the Life Cycle: Evidence for Academic Scientists." *American Economic Review* **81**(1): 114-32.

Levin, Sharon G., Paula E. Stephan, and Mary Beth Walker. 1995. "Planck's Principle Revisited: A Note." *Social Studies of Science* **25**(2): 275-283.

Li, Danielle. 2017. "Expertise vs. Bias in Evaluation: Evidence from the NIH." *American Economic Journal: Applied Economics* **9**(2): 60-92.

Lin, Jimmy, and W. John Wilbur. 2007. "PubMed Related Articles: A Probabilistic Topic-based Model for Content Similarity." *BMC Bioinformatics* **8**(423): 1-14.

Lotka, Alfred J. 1926. "The Frequency Distribution of Scientific Productivity." *Journal of the Washington Academy of Sciences* **16**(12): 317-323.

Luce, R. Duncan, and Albert D. Perry. 1949. "A Method of Matrix Analysis of Group Structure." *Psychometrika* **14**(2): 95-116.

Merton, Robert K. 1973. *The Sociology of Science: Theoretical and Empirical Investigation.* Chicago, IL: University of Chicago Press.

Mohnen, Myra. 2017. "Stars and Brokers: Knowledge Spillovers Among Medical Scientists." Working Paper, Working Paper, University of Essex.

Mokyr, Joel. 1990. *The Lever of Riches: Technological Creativity and Economic Progress.* New York: Oxford University Press.

Mokyr, Joel. 2002. *The Gifts of Athena: Historical Origins of the Knowledge Economy.* Princeton, NJ: Princeton University Press.

Morange, Michel. 1998. *A History of Molecular Biology.* Cambridge, MA: Harvard University Press.

Myers, Kyle. 2018. "The Elasticity of the Direction of Science." Working Paper, National Bureau of Economic Research.

Nagaoka, Sadao, and Hideo Owan. 2014. "Author Ordering in Scientifice Research: Evidence from Scientists Survey in the US and Japan." IIR Working Paper #13-23, Hitotsubashi University, Institute of Innovation Research.

Oettl, Alexander. 2012. "Reconceptualizing Stars: Scientist Helpfulness and Peer Performance." *Management Science* **58**(6): 1122-1140.

Reese, Thomas S. 2004. "My Collaboration with John Heuser." *European Journal of Cell Biology* **83**(6): 243-244.

Romer, Paul M. 1990. "Endogenous Technological Change." *Journal of Political Economy* **98**(5): S71-S102.

Rosvall, Martin, and Carl T. Bergstrom. 2008. "Maps of Random Walks on Complex Networks Reveal Community Structure." *Proceedings of the National Academy of Sciences* **105**(4): 1118-1123.

Sampat, Bhaven N., and Frank R. Lichtenberg. 2011. "What Are the Respective Roles of the Public and Private Sectors in Pharmaceutical Innovation?" *Health Affairs* **30**(2): 332-339.

Santos Silva, J.M.C., and Silvanna Tenreyro. 2006. “The Log of Gravity.” *Review of Economics and Statistics* **88**(4): 641-658.

Shapin, Steven. 1996. *The Scientific Revolution*. Chicago, IL: University of Chicago Press.

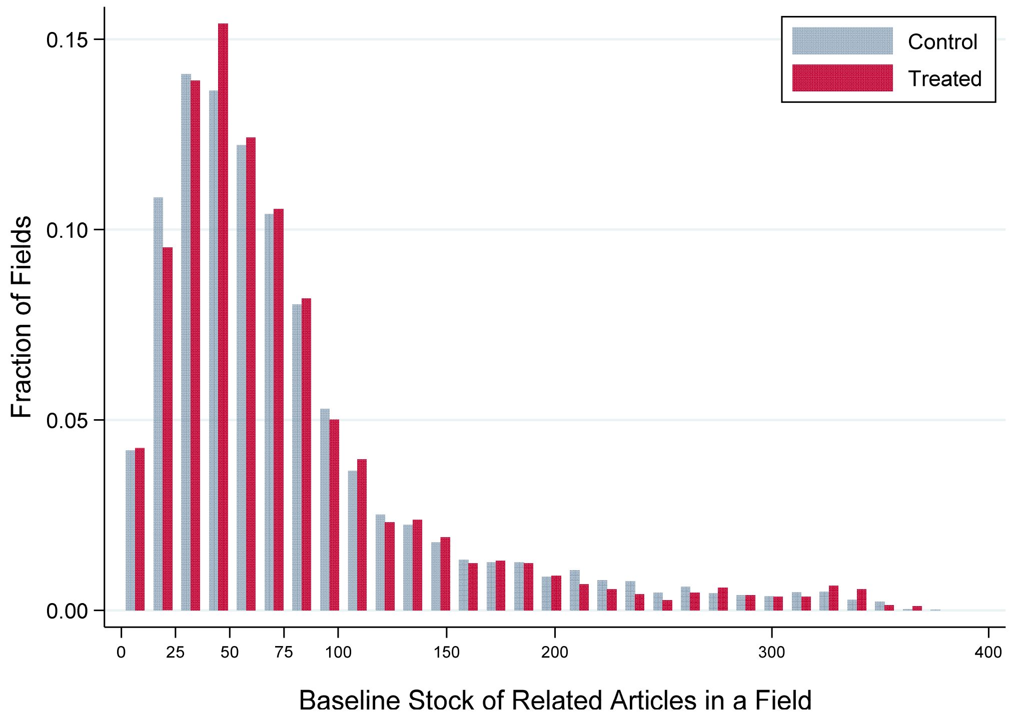
Shwed, Uri, and Peter S. Bearman. 2010. “The Temporal Structure of Scientific Consensus Formation.” *American Sociological Review* **75**(6): 817-840.

Solow, Robert M. 1957. “Technical Change and the Aggregate Production Function.” *Review of Economics and Statistics* **39**(3): 312-320.

Wooldridge, Jeffrey M. 1997. “Quasi-Likelihood Methods for Count Data.” In M. Hashem Pesaran, and Peter Schmidt (Eds.), *Handbook of Applied Econometrics*, pp. 352-406. Oxford: Blackwell.

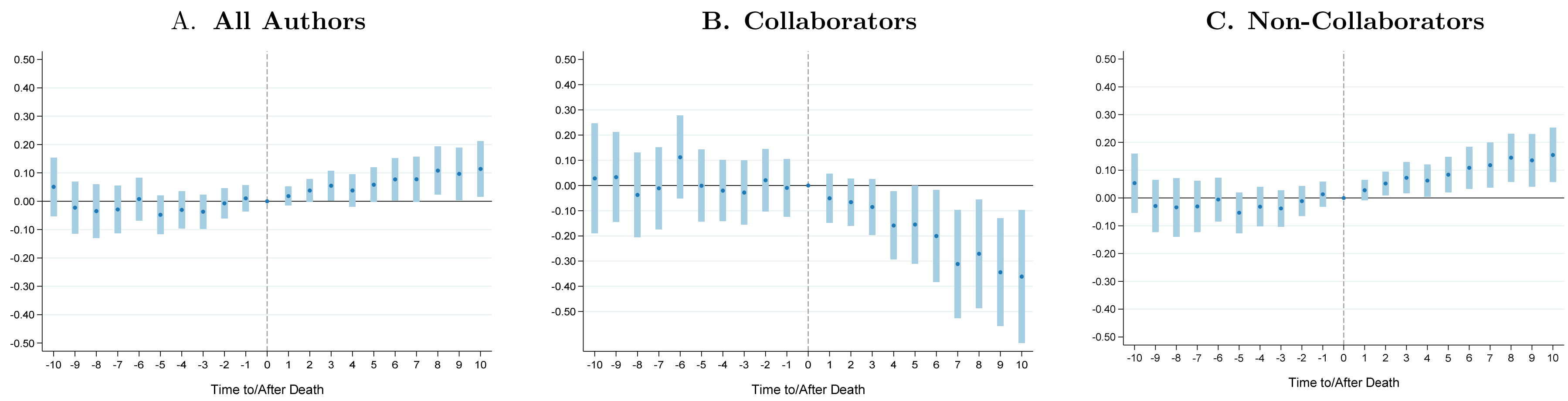
Zuckerman, Harriet A. 1968. “Patterns of Name Ordering Among Authors of Scientific Papers: A Study of Social Symbolism and Its Ambiguity.” *American Journal of Sociology* **74**(3): 276-291.

Figure 1: Cumulative Stock of Publications at Time of Death



Note: We compute the cumulative number of publications, up to the year that immediately precedes the year of death (or counterfactual year of death), between 3,076 treated subfields and 31,142 control subfields.

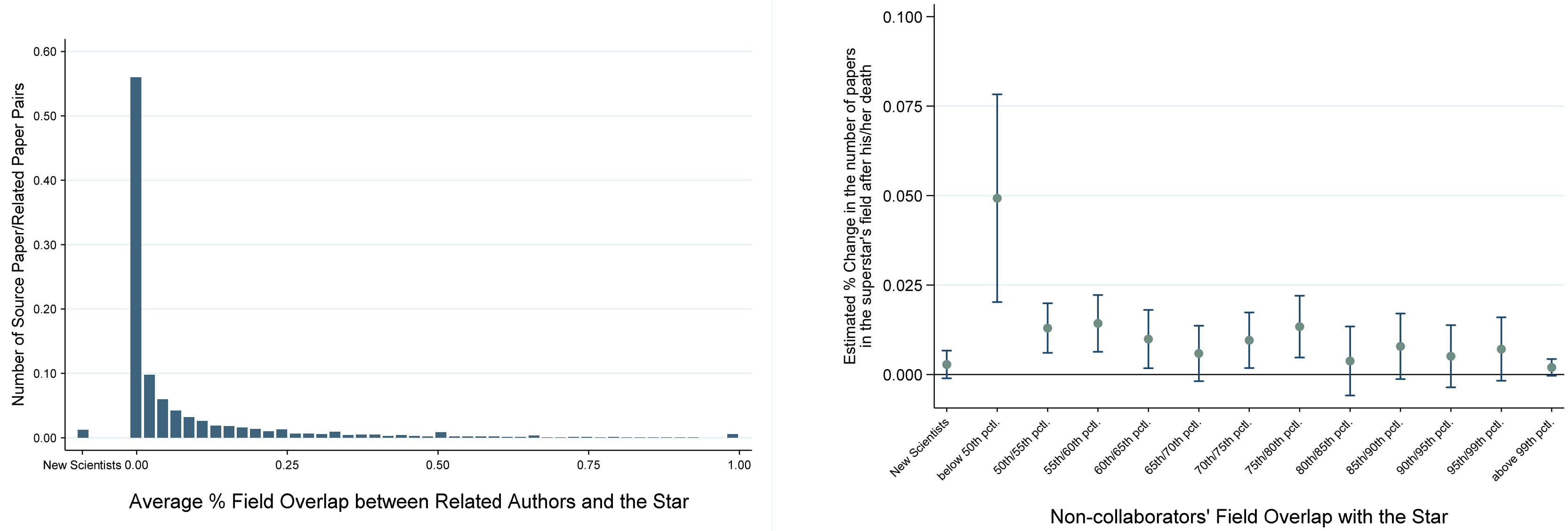
Figure 2
Effect of Star Scientist Death on Subfield Growth and Decline



Note: The dark blue dots in the above plots correspond to coefficient estimates stemming from conditional (subfield) fixed effects Poisson specifications in which publication flows in subfields are regressed onto year effects, subfield age effects, as well as 20 interaction terms between treatment status and the number of years before/after the death event (the indicator variable for treatment status interacted with the year of death is omitted). The specifications also include a full set of lead and lag terms common to both the treated and control subfields to fully account for transitory trends in subfield activity around the time of the death. The 95% confidence interval (corresponding to robust standard errors, clustered around star scientist) around these estimates is plotted with vertical light blue lines; Panel A corresponds to a dynamic version of the specification in column (1) of Table 3; Panel B corresponds to a dynamic version of the specification in column (2) of Table 3; Panel C corresponds to a dynamic version of the specification in column (3) of Table 3.

Figure 3
Characteristics of Related Authors: Competitors or Outsiders?

A. Distribution of Intellectual Proximity ### B. Entering Authors & Field Overlap



Note: Panel A displays the distribution of overlap between the past output of related authors and each star's subfield. For each author on a related article matched to the AAMC Faculty Roster, we create a metric of intellectual proximity by computing the fraction of their publications that belongs to the star's subfield. Slightly more than half of related articles have authors with zero overlap, i.e., this related article is their first contribution to the star's subfield. 1.24% of related articles are authored by new scientists for whom this publication within the subfield is also their first publication overall. Using this information, we aggregate the number of related articles in a particular subfield and in a particular year, e.g., *"the number of articles in the subfield in year t that have authors above the 95th percentile in our measure of field overlap."* In Panel B, each dot corresponds to the magnitude of the treatment effect in a separate regression where the dependent variable is the number of articles in each subfield authored by scientists who belong to a particular bin of intellectual proximity, as measured by field overlap above.

Table 1: Summary Statistics — Deceased Superstar Scientists (N=452)

	Mean	Median	Std. Dev.	Min.	Max.
Year of Birth	1930.157	1930	11.011	1899	1959
Degree Year	1957.633	1957	11.426	1928	1986
Year of Death	1991.128	1992	8.055	1975	2003
Age at Death	60.971	61	9.778	34	91
Female	0.102	0	0.303	0	1
MD Degree	0.403	0	0.491	0	1
PhD Degree	0.489	0	0.500	0	1
MD/PhD Degree	0.108	0	0.311	0	1
Sudden Death	0.409	0	0.492	0	1
Nb. of Subfields	6.794	4	7.305	1	57
Career Nb. of Pubs.	138.221	112	115.704	12	1,380
Career Nb. of Citations	8,341	5,907	8,562	120	72,122
Career NIH Funding	$16,637,919	$10,899,139	$25,441,933	0	$329,968,960
Sits on NIH Study Section	0.007	0	0.081	0	1
Career Nb. of Editorials	0.131	0	0.996	0	17

Note: Sample consists of 452 superstar life scientists who died while still actively engaged in research. See Appendix A for more details on sample construction.

Table 2: Summary Statistics — Control & Treated Subfields at Baseline

	Mean	Median	Std. Dev.	Min.	Max.
Control Subfields (N=31,142)					
Baseline Stock of Related Articles in the Field	76.995	59	64.714	0	384
Baseline Stock of Related Articles in the Field, Non-Collaborators	68.390	51	60.222	0	381
Baseline Stock of Related Articles in the Field, Collaborators	8.604	5	10.358	0	125
Source Article Nb. of Authors	3.970	4	1.901	1	15
Source Article Citations at Baseline	16.331	8	30.305	0	770
Source Article Long-run Citations	70.427	38	116.108	1	4495
Investigator Gender	0.067	0	0.249	0	1
Investigator Year of Degree	1960.546	1962	10.998	1926	1991
Death Year	1991.125	1991	7.968	1975	2003
Age at Death	58.100	58	8.795	34	91
Investigator Cumulative Nb. of Publications	164	131	123	1	1,109
Investigator Cumulative NIH Funding at Baseline	$18,784,517	$11,904,846	$25,160,518	0	$387,558,656
Investigator Cumulative Nb. of Citations	12,141	8,010	12,938	9	157,581
Treated Subfields (N=3,076)					
Baseline Stock of Related Articles in the Field	76.284	58	64.046	0	368
Baseline Stock of Related Articles in the Field, Non-Collaborators	67.752	51	59.725	0	357
Baseline Stock of Related Articles in the Field, Collaborators	8.532	5	9.841	0	86
Source Article Nb. of Authors	3.987	4	1.907	1	14
Source Article Citations at Baseline	16.694	8	36.334	0	920
Source Article Long-run Citations	70.432	35	180.528	1	6598
Investigator Gender	0.099	0	0.299	0	1
Investigator Year of Degree	1960.141	1961	10.898	1928	1986
Death Year	1991.125	1991	7.970	1975	2003
Age at Death	58.100	58	8.796	34	91
Investigator Cumulative Nb. of Publications	170	143	118	12	1,380
Investigator Cumulative NIH Funding at Baseline	$17,637,726	$12,049,690	$24,873,018	0	$329,968,960
Investigator Cumulative Nb. of Citations	11,580	8,726	10,212	120	72,122

Note: The sample consists of subfields for 452 deceased superstar life scientists and their matched control subfields. See Appendix D for details on the matching procedure. All time-varying covariates are measured in the year of superstar death.

Table 3: Effect of Superstar Death on Subfield Entry Rates

	Publication Flows			NIH Funding Flows (Nb. of Awards)		
	All Authors	Collaborators Only	Non-Collaborators Only	All Authors	Collaborators Only	Non-Collaborators Only
	(1)	(2)	(3)	(4)	(5)	(6)
After Death	0.051†	-0.232**	0.082**	0.046	-0.265**	0.110**
	(0.029)	(0.057)	(0.029)	(0.035)	(0.076)	(0.033)
Nb. of Investigators	6,260	6,124	6,260	6,215	5,678	6,202
Nb. of Fields	34,218	33,096	34,218	33,912	29,163	33,806
Nb. of Field-Year Obs.	1,259,176	1,217,905	1,259,176	1,049,942	902,873	1,046,678
Log Likelihood	-2,891,116	-729,521	-2,768,257	-1,350,208	-472,329	-1,223,915

Note: Estimates stem from conditional (subfield) fixed effects Poisson specifications. The dependent variable is the total number of publications in a subfield in a particular year (columns 1, 2, and 3), or the total number of NIH grants that acknowledge a publication in a subfield (columns 4, 5, and 6). All models incorporate a full suite of year effects and subfield age effects, as well as a term common to both treated and control subfields that switches from zero to one after the death of the star, to address the concern that age, year and individual fixed effects may not fully account for trends in subfield entry around the time of death. Exponentiating the coefficients and differencing from one yield numbers interpretable as elasticities. For example, the estimates in column (3) imply that treated subfields see an increase in the number of contributions by non-collaborators after the superstar passes away—a statistically significant $100\times(\exp[0.082]-1)=8.55\%$. The number of observations varies slightly across columns because the conditional fixed effects specification drops observations corresponding to subfields for which there is no variation in activity over the entire observation period.

Robust standard errors in parentheses, clustered at the level of the star scientist. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$.

Table 4: Scientific Impact of Entry

	Vintage-specific long-run citation quantile						
	All Pubs	Bttm. Quartile	2^{nd} Quartile	3^{rd} Quartile	Btw. 75^{th} and 95^{th} pctl.	Btw. 95^{th} and 99^{th} pctl.	Above 99^{th} pctl.
After Death	0.082^{**}	-0.028	0.008	0.031	0.125^{**}	0.232^{**}	0.320^{**}
	(0.029)	(0.036)	(0.033)	(0.032)	(0.035)	(0.049)	(0.081)
Nb. of Investigators	6,260	6,222	6,260	6,257	6,255	6,161	5,283
Nb. of Fields	34,218	33,714	34,206	34,212	34,210	33,207	21,852
Nb. of Field-Year Obs.	1,259,176	1,240,802	1,258,738	1,258,954	1,258,880	1,221,952	804,122
Log Likelihood	-2,768,257	-689,467	-1,125,554	-1,432,227	-1,469,094	-542,731	-156,519

Note: Estimates stem from conditional (subfield) fixed effects Poisson specifications. The dependent variable is the total number of publications by non-collaborators in a subfield in a particular year, where these publications fall in a particular quantile bin of the long-run, vintage-adjusted citation distribution for the universe of journal articles in *PubMed*. All models incorporate a full suite of year effects and subfield age effects, as well as a term common to both treated and control subfields that switches from zero to one after the death of the star. Exponentiating the coefficients and differencing from one yield numbers interpretable as elasticities. For example, the estimates in column (1), Panel A, imply that treated subfields see an increase in the number of contributions by non-collaborators after the superstar passes away—a statistically significant 100×(exp[0.082]-1)=8.55%.

Robust standard errors in parentheses, clustered at the level of the star scientist. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$.

Table 5: Entry and Research Direction

Panel A	Cardinal Measure		Ordinal Measure	
	Intllct. Proximate Articles	Intllct. Distant Articles	Intllct. Proximate Articles	Intllct. Distant Articles
After Death	0.091**	0.028	0.117**	-0.024
	(0.030)	(0.035)	(0.028)	(0.037)
Nb. of Investigators	6,228	6,099	6,260	6,017
Nb. of Fields	33,375	32,232	34,218	31,712
Nb. of Field-Year Obs.	1,228,157	1,186,589	1,259,176	1,167,423
Log Likelihood	-1,628,374	-1,816,449	-1,893,982	-1,628,170
Panel B	**In-field vs. Out-of-field References**		**Backward Citations to the Star's Bibliome**	
	w/ in-field references	w/o in-field references	w/ references to the star	w/o references to the star
After Death	-0.023	0.128**	0.078*	0.152**
	(0.041)	(0.031)	(0.036)	(0.034)
Nb. of Investigators	6,195	6,260	6,247	6,259
Nb. of Fields	32,721	34,218	34,179	34,147
Nb. of Field-Year Obs.	1,204,315	1,259,176	1,257,747	1,256,576
Log Likelihood	-792,803	-2,510,350	-1,914,447	-1,767,579
Panel C	**Vintage of Cited References**		**Vintage of 2-way MeSH term combinations**	
	Young	Old	Young	Old
After Death	0.071*	-0.010	0.090**	0.029
	(0.035)	(0.034)	(0.033)	(0.036)
Nb. of Investigators	6,260	6,260	6,258	6,260
Nb. of Fields	34,218	34,214	34,206	34,210
Nb. of Field-Year Obs.	1,259,176	1,259,044	1,258,732	1,258,906
Log Likelihood	-2,124,598	-1,613,454	-1,853,064	-1,784,279

Note: Estimates stem from conditional (subfield) fixed effects Poisson specifications. In Panel A, the dependent variable is the total number of publications by non-collaborators in a subfield in a particular year, where these publications can either be proximate in intellectual space to the star's source publication, or more distant (in the PMRA sense). Since PMRA generates both a cardinal and an ordinal measure of intellectual proximity, we parse the related articles using both measures, yielding a total of four different specifications. For the cardinal measure, a related article is deemed proximate if its similarity score is above .58, which corresponds to the median of relatedness in the sample. For the ordinal measure, a related article is deemed proximate if its similarity rank is below 90, which also corresponds to the median of similarity in the sample. In Panel B, we focus on whether the content of entrants' contributions in the subfield change after the superstar passes away. Each cited reference in a related article can either belong to the subfield, or fall outside of it; it can cite a publication of the star scientist associated with the subfield, or fail to cite any of the star's past contributions. In Panel C, the dependent variable is the total number of publications by non-collaborators in a subfield in a particular year, where these publications can either be "fresh" (citing young references, or being annotated by MeSH terms of recent vintage) or stale (citing old references, or being annotated by MeSH terms of distant vintage). All models incorporate a full suite of year effects and subfield age effects, as well as a term common to both treated and control subfields that switches from zero to one after the death of the star. Exponentiating the coefficients and differencing from one yield numbers interpretable as elasticities. For example, the estimates in the first column of Panel A imply that treated subfields see an increase in the number of PMRA-proximate contributions by non-collaborators after the superstar passes away—a statistically significant 100×(exp[0.091]-1)=9.53%. Robust standard errors in parentheses, clustered at the level of the star scientist. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$.

Table 6: Breakdown by Star Scientist Characteristics

	Publications		Citations		Funding		Importance to the Field	
	Below Median	Above Median	Below Median	Above Median	Below Median	Above Median	Below Median	Above Median
After Death	0.059	0.116^{*}	0.036	0.125^{**}	0.014	0.162^{**}	0.063^{*}	0.123^{**}
	(0.037)	(0.050)	(0.042)	(0.040)	(0.040)	(0.052)	(0.031)	(0.045)
Nb. of Investigators	2,901	4,836	2,792	4,619	3,048	4,287	5,019	4,493
Nb. of Fields	17,210	17,008	17,328	16,890	15,731	15,487	16,985	17,233
Nb. of Field-Year Obs.	632,089	627,087	636,750	622,426	578,277	570,665	625,140	634,036
Log Likelihood	-1,377,741	-1,387,648	-1,367,337	-1,396,654	-1,268,567	-1,252,952	-1,462,541	-1,257,972

Note: Estimates stem from conditional (subfield) fixed effects Poisson specifications. The dependent variable is the total number of publications by non-collaborators in a subfield in a particular year. Each pair of columns splits the sample across the median of a particular covariate for the sample of fields (treated and control) in the baseline year. The table examines differences in the extent to which the eminence of the star at death (respectively counterfactual year of death for controls) influences the rate at which non-collaborators enter the field after the star passes away. Eminence is measured through the star's cumulative number of publications, the star's cumulative number of citations garnered up to the year of death, and the star's cumulative amount of NIH funding. We also have a "local" measure of eminence: the star's importance to the field, which is defined as the proportion of articles in the subfield up to the year of death for which the star is an author. All models incorporate a full suite of year effects and subfield age effects, as well as a term common to both treated and control subfields that switches from zero to one after the death of the star. Exponentiating the coefficients and differencing from one yield numbers interpretable as elasticities. For example, the estimate in the second column implies that treated subfields see an increase in the number of contributions by non-collaborators after the superstar passes away—a statistically significant 100×(exp[0.116]-1)=12.30%.

Robust standard errors in parentheses, clustered at the level of the star scientist. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$.

Table 7: The Nature of Entry Barriers

Panel A	Subfield Coherence					
	PMRA-based definition		Citation-based definition		Cliquishness	
	Below Median	Above Median	Below Median	Above Median	Below Median	Above Median
After Death	0.202**	0.067	0.161**	0.096*	0.129**	0.064
	(0.038)	(0.048)	(0.053)	(0.041)	(0.049)	(0.052)
Nb. of Investigators	3,353	3,203	3,422	3,157	2,865	3,561
Nb. of Fields	9,062	7,828	8,731	8,159	8,044	8,846
Nb. of Field-Year Obs.	334,142	288,284	321,826	300,600	296,704	325,722
Log Likelihood	-711,335	-664,170	-760,842	-631,287	-692,330	-685,682

Panel B	Indirect Control through Collaborators					
	Editorial Channel		NIH Study Section Channel		Fraction of Subfield NIH Funding	
	Below Median	Above Median	Below Median	Above Median	Below Median	Above Median
After Death	0.147**	0.086†	0.134**	-0.078	0.174**	0.084
	(0.056)	(0.048)	(0.043)	(0.095)	(0.051)	(0.051)
Nb. of Investigators	3,452	2,068	4,385	664	3,558	2,526
Nb. of Fields	11,110	5,780	15,338	1,552	9,860	7,030
Nb. of Field-Year Obs.	410,025	212,401	565,219	57,207	363,584	258,842
Log Likelihood	-951,705	-461,769	-1,293,997	-125,950	-840,666	-545,869

Note: Estimates stem from conditional (subfield) fixed effects Poisson specifications. The dependent variable is the total number of publications by non-collaborators in a subfield in a particular year. The sample is limited to the subfields in which the most eminent among the stars were active (specifically, above the median of the "cumulative citations up to the year of death" metric). Each pair of columns splits the sample across the median of a particular covariate for the sample of subfields (treated and control) in the baseline year. For example, the first two columns of Panel B compare the magnitude of the treatment effect for stars whose collaborators have written an above-median number of editorials in the five years preceding the superstar's death, vs. a below-median number of editorials. All models incorporate a full suite of year effects and subfield age effects, as well as a term common to both treated and control subfields that switches from zero to one after the death of the star. Exponentiating the coefficients and differencing from one yield numbers interpretable as elasticities. For example, the estimates in the first column of Panel B imply that treated subfields see an increase in the number of contributions by non-collaborators after the superstar passes away—a statistically significant 100×(exp[0.147]-1)=15.84%.

Robust standard errors in parentheses, clustered at the level of the star scientist. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$.

Table 8: The Eminence of Entrants—The Circle of Life

	Star Related Author		Elite Related Author		
	No	Yes	Never	Current	Future
After Death	0.103**	0.055†	0.066*	0.077	0.205**
	(0.036)	(0.030)	(0.029)	(0.052)	(0.074)
Nb. of Investigators	6,254	6,260	6,260	5,721	5,886
Nb. of Fields	34,160	34,218	34,218	28,992	29,650
Nb. of Field-Year Obs.	1,257,053	1,259,176	1,259,176	1,067,107	1,091,439
Log Likelihood	-1,287,272	-2,324,369	-2,615,424	-373,036	-377,540

Note: Estimates stem from conditional (subfield) fixed effects Poisson specifications. The dependent variable is the total number of publications by non-collaborators in a subfield in a particular year, where these publications have scientists on their authorship roster with certain demographic characteristics. The first two columns examine the differential effect of the publications in the subfield having a star author vs. no star author. We rely on our home-grown definition of star—a fixed universe of 12,935 individuals that are in some sense "born" as stars. In the next two columns, we focus on two of our metrics of stardom: becoming a Howard Hughes Medical Investigator and or becoming a member of the National Academy of Science/Medicine. At a given point of time, every related author either (i) is already a member of this rarefied elite; (ii) will be member of it in the future; or (iii) will never become a member of it, and this taxonomy provides a basis to split the output of each subfield into three non-overlapping categories in each year. All models incorporate a full suite of year effects and subfield age effects, as well as a term common to both treated and control subfields that switches from zero to one after the death of the star. Exponentiating the coefficients and differencing from one yield numbers interpretable as elasticities. For example, the estimates in the first column imply that treated subfields see an increase in the number of contributions by non-stars after the superstar passes away—100×(exp[0.103]-1)=10.85%.

Robust standard errors in parentheses, clustered at the level of the star scientist. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$.

Supplementary Online Material

Appendix A:
Criteria for Delineating the Set of 12,935 "Superstars"

Highly Funded Scientists. Our first data source is the Consolidated Grant/Applicant File (CGAF) from the U.S. National Institutes of Health (NIH). This dataset records information about grants awarded to extramural researchers funded by the NIH since 1938. Using the CGAF and focusing only on direct costs associated with research grants, we compute individual cumulative totals for the decades 1977-1986, 1987-1996, and 1997-2006, deflating the earlier years by the Biomedical Research Producer Price Index. We also recompute these totals excluding large center grants that usually fund groups of investigators (M01 and P01 grants). Scientists whose totals lie above the 95^{th} percentile of e'ither distribution constitute our first group of superstars. In this group, the least well-funded investigator garnered \$10.5 million in career NIH funding and the most well-funded \$462.6 million.[i]

Highly Cited Scientists. Despite the preeminent role of the NIH in the funding of public biomedical research, the above indicator of "superstardom" biases the sample towards scientists conducting relatively expensive research. We complement this first group with a second composed of highly cited scientists identified by the Institute for Scientific Information. A Highly Cited listing means that an individual was among the 250 most cited researchers for their published articles between 1981 and 1999, within a broad scientific field.[ii]

Top Patenters. We add to these groups academic life scientists who belong in the top percentile of the patent distribution among academics—those who were granted 17 patents or more between 1976 and 2004.

Members of the National Academy of Science and of the Institute of Medicine. We add to these groups academic life scientists who were elected to the National Academy of Science or the Institute of Medicine between 1970 and 2013.

MERIT Awardees of the NIH. Initiated in the mid-1980s, the MERIT Award program extends funding for up to 5 years (but typically 3 years) to a select number of NIH-funded investigators *"who have demonstrated superior competence, outstanding productivity during their previous research endeavors and are leaders in their field with paradigm-shifting ideas."* The specific details governing selection vary across the component institutes of the NIH, but the essential feature of the program is that only researchers holding an R01 grant in its second or later cycle are eligible. Further, the application must be scored in the top percentile in a given funding cycle. We add to this category the NIH Director's Pioneer Awardees. Part of the "High-Risk, High-Reward Research" program, since 2004 the award has supported "scientists with outstanding records of creativity pursuing new research directions to develop pioneering approaches to major challenges in biomedical and behavioral research."

Former and current Howard Hughes Medical Investigators (HHMIs). Every three years, the Howard Hughes Medical Institute selects a small cohort of mid-career biomedical scientists with the potential to revolutionize their respective subfields. Once selected, HHMIs continue to be based at their institutions,

[i] We perform a similar exercise for scientists employed by the intramural campus of the NIH. These scientists are not eligible to receive extramural funds, but the NIH keeps records of the number of "internal projects" each intramural scientist leads. We include in the elite sample the top five percentiles of intramural scientists according to this metric.

[ii] The relevant scientific fields in the life sciences are microbiology, biochemistry, psychiatry/psychology, neuroscience, molecular biology & genetics, immunology, pharmacology, and clinical medicine.

typically leading a research group of 10 to 25 students, postdoctoral associates and technicians. Their appointment is reviewed every five years, based solely on their most important contributions during the cycle.[iii]

Early career prize winners. We also included winners of the Pew, Searle, Beckman, Rita Allen, and Packard scholarships for the years 1981 through 2000. Every year, these charitable foundations provide seed funding to between 20 and 40 young academic life scientists. These scholarships are the most prestigious accolades that young researchers can receive in the first two years of their careers as independent investigators.

Consolidated categories. Why use 8 different criteria to delineate the set of stars? There are two reasons to do so. First, there is of course no agreed-upon definition of stardom in academic science, and choosing an eclectic set of metric makes it less likely that our analysis will be biased by the idiosyncrasies of any particular metric. For example, the funding metric will tend to bias the set of stars towards scientists doing relatively expensive research (e.g., clinical research, or research on monkeys/other mammals vs. research on invertebrates such as the nematode worm *c. elegans*). Table A1 documents the overlap between each of the eight metrics. Some metrics are highly negatively correlated (e.g., ECPW and high NIH funding) while most correlations between individual metrics are modest in magnitude.

Second, if we focused on a single, incontrovertible metric such as election to the National Academy of Sciences, we would not have enough statistical power to identify the main effect of death on subfield growth. To examine the effect of star death across stars of different types, we consolidate the eight metrics into three mutually exclusive categories:

(i) "Cumulative stars," who enter the sample on the basis of cumulative achievement (high NIH grant receipt, highly cited scientists, top patenters, and members of the National Academy of Science/Medicine ($N = 6,858$ or 53%);

(ii) "Shooting stars," who enter the sample on the basis of a specific contribution (appointment as a Howard Hughes Medical Investigator; NIH MERIT/Director Pioneer awardees; Early career prize winners), with no presumption that this mark of elevated status will endure over the entire career ($N = 3,859$ or 30%);

(iii) "Cumulative$\oplus$Shooting stars," who enter the sample based on at least one cumulative metric, and at least one "burst" metric ($N = 2,218$ or 17%).

We also create a subsample limited to the members of the National Academies of Science/Medicine and Investigators of the Howard Hughes Medical Institute. One can think of this rarefied subset (which *de facto* subsumes Nobel prize winners and Lasker awardees) as "the elite within the elite" of academic biomedical research (N=3,325 or 26% of the total).

In Table A2, we run our benchmark specification (number of papers in the field by non-collaborators, as in the third column of Table 3) separately on these four subsamples. All the coefficients are positive in magnitude, but some of them are imprecisely estimated. Table A3 lists all of the 452 extinct stars in the sample, along with basic demographic information, cause of death, institutional affiliation, and a short description of their research expertise.

[iii] See Azoulay et al. (2011) for more details and an evaluation of this program.

Table A1: Star Decomposition

	Highly Funded	Highly Cited	Top Patenter	NAS	NAM	MERIT	HHMI	ECPW
Highly Funded	7,822	886	189	942	1,033	1,540	221	128
Highly Cited	886	1,921	96	385	355	442	141	58
Top Patenter	189	96	606	88	55	86	29	14
NAS	942	385	88	1,843	430	561	295	151
NAM	1,033	355	55	430	1,933	368	176	68
MERIT	1,540	442	86	561	368	2,898	196	145
HHMI	221	141	29	295	176	196	866	179
ECPW	128	58	14	151	68	145	179	1,114

Note: Metrics of stardom and their distribution in the sample of 12,935 eminent scientists. NAS=National Academy of Sciences; NAM=National Academy of Medicine; MERIT=Method to Extend Research In Time, an exceptional NIH grant category; HHMI=Howard Hughes Medical Investigator; ECPW=Early Career Prize Winners.

Table A2: Impacts by Type of Star

	Shooting Stars	Cumulative Stars	Shooting & Cumulative Stars	"Elite of the Elite"
After Death	0.047	0.079*	0.154*	0.032
	(0.056)	(0.038)	(0.069)	(0.052)
Nb. of Investigators	1,551	3,164	1,545	1,708
Nb. of Fields	6,584	16,095	11,539	11,855
Nb. of Field-Year Obs.	242,409	592,030	424,737	436,081
Log Likelihood	-535,715	-1,345,402	-938,102	-952,496

Note: Estimates stem from conditional (subfield) fixed effects Poisson specifications. The dependent variable is the total number of publications by non-collaborators in a subfield in a particular year, contributed by non-collaborators. All models incorporate a full suite of year effects and subfield age effects, as well as a term common to both treated and control subfields that switches from zero to one after the death of the star, to address the concern that age, year and individual fixed effects may not fully account for trends in subfield entry around the time of death for the deceased star. Exponentiating the coefficients and differencing from one yield numbers interpretable as elasticities.

Robust standard errors in parentheses, clustered at the level of the star scientist.
$^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$.

Table A3: List of 452 Extinct Superstars

Investigator Name			Cause of death if known	Institution at the time of death	Scientific domain
Richard C. Parker	[1952-1986]	PhD, 1979	lymphoma	Columbia University	properties of cellular and viral src genes
Richard E. Weitzman	[1943-1980]	MD, 1968	cancer	Harbor-UCLA Medical Center	arginine vasopressin metabolism
Eva U.J. Paucha	[1949-1988]	PhD, 1976	cancer	Dana Farber Cancer Institute	mechanism of transformation by SV40 large T antigen
Kiertisin Dharmsathaphorn	[1950-1990]	MD, 1972	AIDS	University of California — San Diego	intestinal secretory mechanisms and antidiarrheal drugs
Ernest G. Peralta	[1959-1999]	PhD, 1986	brain cancer	Harvard University	signal transduction mechanisms of muscarinic receptors
Roderich Walter	[1937-1979]	PhD, 1964	malignant melanoma	University of Illinois	solid-phase peptide synthesis
JoAnn E. Franck	[1950-1992]	PhD, 1981	cancer	University of Washington School of Medicine	hippocampal damage as a cause of epilepsy
Thomas K. Tatemichi	[1952-1995]	MD, 1978	non hodgkin's lymphoma	Columbia University College of Physicians & Surgeons	mechanisms and syndromes of dementia related to stroke
Bruce S. Schoenberg	[1942-1987]	MD, 1968	cancer	NIH	prevention and control of neurological disorders
George Khoury	[1943-1987]	MD, 1970	lymphoma	NIH	genetics of simian virus 40, human papovavirus and HIV
Leonard N. Horowitz	[1947-1992]	MD, 1972	cancer	University of Pennsylvania School of Medicine	diagnosing and treatment of ventricular arrythmia
W. Alden Spencer	[1931-1977]	MD, 1956	long illness	Columbia University	plasticity of the simplest neuronal pathways
Jerome T. Pearlman	[1933-1979]	MD, 1957	prolonged illness	UCLA	laboratory studies of retinal degenerations
Joram Heller	[1934-1980]	MD/PhD, 1965	brain cancer	UCLA	biochemical and biophysical investigation of rhodopsin
B. Frank Polk	[1942-1988]	MD, 1967	brain cancer	Johns Hopkins University School of Medicine	epidemiology of HIV infection
Ronald D. Fairshter	[1942-1988]	MD, 1968	rapidly metastatic melanoma	University of California — Irvine	clinical studies in chronic obstructive pulmonary disease
Cornelia P. Channing	[1938-1985]	PhD, 1966	breast cancer	University of Maryland School of Medicine	mechanism of luteinization in vitro and in vivo
Joel D. Meyers	[1944-1991]	MD, 1970	colon cancer	University of Washington/FHCRC	infections caused by suppression of the immune system in organ transplant and AIDS patients
Richard L. Lyman	[1927-1975]	PhD, 1957	terminal illness for months	University of California — Berkeley	protein, trypsin inhibitors and pancreatic secretion
James N. Gilliam	[1936-1984]	MD, 1964	cancer	University of Texas Southwestern Medical Center at Dallas	cutaneous lupus erythematosus pathogenesis mechanisms
Gordon M. Tomkins	[1926-1975]	MD/PhD, 1953	brain surgery to remove a tumor	University of California — San Francisco	pleiotypic response in regulation of cell growth
Muriel R. Steele	[1930-1979]	MD, 1957	metastatic disease	University of California — San Francisco	surgical treatment of liver trauma
Allastair M. Karmody	[1937-1986]	MD, 1963	gastric cancer	Albany Medical College	novel procedures for difficult vascular surgical problems
Chaviva Isersky	[1937-1986]	PhD, 1967	cancer	NIH/NIDDK	Characterization of the protein responsible for amyloidosis
Melvin L. Marcus	[1940-1989]	MD, 1966	colon cancer	UMASS	cardiology, heart disease, coronary vascular adaptations to myocardial hypertrophy
Alan S. Morrison	[1943-1992]	PhD, 1972	cancer	Brown University Medical School	hormones in the epidemiology of prostatic hyperplasia
Sidney Futterman	[1929-1979]	PhD, 1954	prolonged illness	University of Washington School of Medicine	biochemistry of the retina and pigment epithelium
Loretta L. Leive	[1936-1986]	PhD, 1963	cancer	NIH/NIDDK	role of bacterial cell surface in microbial physiology and pathogenesis
Philip G. Weiler	[1941-1991]	MD, 1965	terminal illness	University of California — Davis	coronary heart disease & stroke in the elderly
Ira M. Goldstein	[1942-1992]	MD, 1966	metastatic lung cancer	University of California — San Francisco	pancreatitis, complement and lung injury
Harold Weintraub	[1945-1995]	MD/PhD, 1973	brain cancer	University of Washington/FHCRC	characterization and function of MyoD gene
Richard K. Gershon	[1932-1983]	MD, 1959	lung cancer	Yale University	immunologic responses to tumor grafts
Edward J. Sachar	[1933-1984]	MD, 1956	stroke three years ago	Columbia University	psychoendocrine studies of schizophrenic reactions
Catherine Cole-Beuglet	[1936-1987]	MD, 1962	colon cancer	University of California — Irvine	ultrasonography of the breast
Theodore S. Zimmerman	[1937-1988]	MD, 1963	lung cancer	Scripps Research Institute	platelet/plasma protein interaction in blood coagulation
Markku Linnoila	[1947-1998]	MD/PhD, 1974	cancer	NIH	studies on the biological bases of impulsivity and aggression
William J. Mellman	[1928-1980]	MD, 1952	lymphoma	University of Pennsylvania School of Medicine	human genetics and pediatrics
Dennis Slone	[1930-1982]	MD, 1956	long illness	Boston University School of Medicine	intensive inpatient psychiatric monitoring program
Roger O. Eckert	[1934-1986]	PhD, 1960	melanoma	UCLA	ionic and metabolic mechanisms in neuronal excitability
Michael Solursh	[1942-1994]	PhD, 1968	AIDS	University of Iowa School of Medicine	extracellular matrix and cell migration
Larry C. Clark	[1948-2000]	PhD, 1981	prostate cancer	University of Arizona	nutritional prevention of cancer
Robert F. Spencer	[1949-2001]	PhD, 1974	gastric carcinoma	Medical College of Virginia	neuroanatomy of the oculomotor system
Carl C. Levy	[1928-1981]	PhD, 1957	leukemia	NIH/NCI	regulation of intracellular messenger RNA
Marshall H. Becker	[1940-1993]	PhD, 1968	intractable illness	University of Michigan, Ann Arbor	elaboration of the health belief model
Samuel W. Perry, 3rd	[1941-1994]	MD, 1967	pancreatic cancer	Cornell University — Weill Medical College	psychological course of prolonged infection among AIDS patients
Michael A. Kirschenbaum	[1944-1997]	MD, 1969	long illness	University of California — Irvine	prostaglandins and kidney medicine
Janis V. Giorgi	[1947-2000]	PhD, 1977	uterine cancer	UCLA	cellular immunology of resistance to HIV
Herbert F. Hasenclever	[1924-1978]	PhD, 1953	cancer	NIH/NIAID	mannan polysaccharides of pathogenic fungi
Edward C. Franklin	[1928-1982]	MD, 1950	brain cancer	New York University School of Medicine	structure and properties of rheumatoid antibodies
Robert M. Joy	[1941-1995]	PhD, 1969	cancer	University of California — Davis	pesticide induced changes in central nervous function
Lois K. Miller	[1945-1999]	PhD, 1972	melanoma	University of Georgia	genetics and molecular biology of baculoviruses
Gerald T. Babcock	[1946-2000]	PhD, 1973	cancer	Michigan State University	bioenergetic mechanisms in multicenter enzymes
John G. Gambertoglio	[1947-2001]	PharmD, 1972	multiple sclerosis	University of California — San Francisco	pharmacokinetics in healthy volunteers and subjects with renal insufficiency and on hemodialysis
John C. Cassel	[1921-1976]	MD, 1946		University of North Carolina at Chapel Hill	Contribution of the social environment to host resistance
Ernst A. Noltmann	[1931-1986]	MD, 1956	severe health problems	University of California — Riverside	biochemical and physical characterization of phosphoglucose isomerase
Edward A. Smuckler	[1931-1986]	MD/PhD, 1963	barrett's disease/oesophagal cancer	University of California — San Francisco	cytochemical studies in liver injury
Joseph W. St. Geme, Jr.	[1931-1986]	MD, 1956	cardiac myopathy	University of Colorado Health Sciences Center	studies of cellular resistance to virus infection
Edwin H. Beachey	[1934-1989]	MD, 1962	cancer	University of Tennessee	chemistry and immunology of streptococcal m proteins
Ora M. Rosen	[1935-1990]	MD, 1960	breast cancer	Sloan Kettering Institute for Cancer Research	Cloning and characterization of gene for human insulin receptor
Tai-Shun Lin	[1939-1994]	PhD, 1970	non hodgkin's lymphoma	Yale University	synthesis and development of nucleoside analogs as antiviral and anticancer compounds
Judith G. Pool	[1919-1975]	PhD, 1946	brain tumor	Stanford University	pathophysiology of hemophilia
Ardie Lubin	[1920-1976]	PhD, 1951	serious illness for months	Naval Health Research Center	repeated measurement design in psychopharmacology
William H. Hildemann	[1927-1983]	PhD, 1956	amyotrophic lateral sclerosis	UCLA	mechanisms of immunoblocking versus tumor immunity
Murray Rabinowitz	[1927-1983]	MD, 1950	muscular dystrophy	University of Chicago	mitochondrial assembly and replication
Paul A. Obrist	[1931-1987]	PhD, 1958	3 year illness	University of North Carolina at Chapel Hill	blood pressure control: relation to behavioral stress
C. Richard Taylor	[1939-1995]	PhD, 1963	heart failure	Harvard University	locomotion–idling metabolism and gait dynamics
Helene S. Smith	[1941-1997]	PhD, 1967	breast cancer	University of California — San Francisco	malignant progression of the human breast/predictors of breast cancer prognosis
Bruce W. Erickson	[1942-1998]	PhD, 1970	cancer	University of North Carolina at Chapel Hill	engineering of nongenetic beta proteins
Norton B. Gilula	[1944-2000]	PhD, 1971	lymphoma	Scripps Research Institute	cell junction biosynthesis and biogenesis/cell-cell communication
John M. Eisenberg	[1946-2002]	MD, 1972	high-grade malignant glioma	Georgetown University Medical Center	health services research
Elizabeth A. Bates	[1947-2003]	PhD, 1974	pancreatic cancer	University of California — San Diego	cross-linguistic studies of language development, processing and breakdown in aphasia
Ira Herskowitz	[1946-2003]	PhD, 1971	pancreatic cancer	University of California — San Francisco	genetics of yeast mating type
Wallace P. Rowe	[1926-1983]	MD, 1948	colon cancer	NIH	genetic basis of disease in murine leukemia viruses
J. Weldon Bellville	[1926-1983]	MD, 1952	cancer	UCLA	dynamic isolation studies of control of respiration
Peter W. Lampert	[1929-1986]	MD, 1955	lymphoma	University of California — San Diego	pathogenesis of virus-induced brain disease
Sheldon D. Murphy	[1933-1990]	PhD, 1958	cancer	University of Washington School of Medicine	biochemical and physiologic response to toxic stress
Allan C. Wilson	[1934-1991]	PhD, 1961	leukemia	University of California — Berkeley	use of molecular approaches to understand evolutionary change
Bernard N. Fields	[1938-1995]	MD, 1962	pancreatic cancer	Harvard Medical School/Brigham & Women's Hospital	genetic and molecular basis of viral injury to the nervous system
Priscilla A. Campbell	[1940-1998]	PhD, 1968	cervical cancer	University of Colorado Health Sciences Center/Natl. Jewish Center	cell biology of the immune response to bacteria
Ethan R. Nadel	[1941-1998]	PhD, 1969	cancer	Yale University	thermoregulation during exercise and heat exposure
Peter A. Kollman	[1944-2001]	PhD, 1970	cancer	University of California — San Francisco	free energy perturbation calculations and their application to macromolecules

Investigator Name			Cause of death if known	Institution at the time of death	Scientific domain
David Tapper	[1945-2002]	MD, 1970	long battle with renal cell carcinoma	University of Washington School of Medicine	determination of a new growth factor in breast milk
Cyril S. Stulberg	[1919-1977]	PhD, 1947	multiple sclerosis	Wayne State University School of Medicine	characterization and preservation of cell strains
Dorothy T. Krieger	[1927-1985]	MD, 1949	breast cancer	Mount Sinai School of Medicine	CNS-pituitary-adrenal interactions
Aaron Janoff	[1930-1988]	PhD, 1959	long illness	SUNY HSC at Stony Brook	pathology of smoking and emphysema
Wylie J. Dodds	[1934-1992]	MD, 1960	brain cancer	Medical College of Wisconsin	esophageal motor function in health and disease
Oscar A. Kletzky	[1936-1994]	MD, 1961	lung cancer	UCLA	ameliorating effects of estrogen replacement therapy on cerebral blood flow and sleep
Nelson Butters	[1937-1995]	PhD, 1964	Lou Gehrig's disease	University of California — San Diego	cognitive deficits related to chronic alcoholism
Elizabeth M. Smith	[1939-1997]	PhD, 1978	cancer	Washington University in St. Louis	psychiatric problems among disaster survivors
David G. Marsh	[1940-1998]	PhD, 1964	glioblastoma	Johns Hopkins University School of Medicine	genetics of allergy and asthma
George C. Cotzias	[1918-1977]	MD, 1944	lung cancer	Cornell University Medical College	studies of extrapyramidal & related behavioral disorders
Robert D. Allen	[1927-1986]	PhD, 1953	pancreatic cancer	Dartmouth Medical School	cytoplasmic rheology of motile cells
Marilyn Bergner	[1933-1992]	PhD, 1970	ovarian cancer	Johns Hopkins University School of Public Health	cost and efficacy of home care for COPD patients
G. Harrison Echols, Jr.	[1933-1993]	PhD, 1959	lung cancer	University of California — Berkeley	Genetic and chemical studies of phage lambda development
Milton H. Stetson	[1943-2002]	PhD, 1970	prolonged and courageous fight with illness	University of Delaware	environmental regulation of reproduction and the onset of puberty
Nicholas R. DiLuzio	[1926-1986]	PhD, 1954	extended illness	Tulane University School of Medicine	role recognition factors and macrophages in neoplasia
Lauran D. Harris	[1927-1987]	MD, 1947	long illness	Boston University School of Medicine	sphincter strength–its measurement and control
Charles W. Mays	[1930-1990]	PhD, 1958	cancer	National Cancer Institute	reducing cancer risk by radionuclide chelation
Lawrence H. Piette	[1932-1992]	PhD, 1957	cancer	Utah State University	electron spin resonance spectroscopy
Mehdi Tavassoli	[1933-1993]	MD, 1961	heart failure	University of Mississippi Medical Center	hematopoietic stem cell purification and biology
Howard M. Temin	[1934-1994]	PhD, 1959	lung cancer	University of Wisconsin	molecular biology and genetics of tumor viruses
Mette Strand	[1937-1997]	PhD, 1964	cancer	Johns Hopkins University School of Medicine	parasite immunochemistry and vaccine development
William L. Chick	[1938-1998]	MD, 1963	diabetes complications	UMASS	studies of islet and beta cells in pancreatic transplantation
Robert A. Mendelson, Jr.	[1941-2001]	PhD, 1968	lung cancer	University of California — San Francisco	molecular mechanism of muscle contraction
Susan M. Sieber	[1942-2002]	PhD, 1971	breast cancer	National Cancer Institute	biochemical epidemiology and cancer
Joachim G. Liehr	[1942-2003]	PhD, 1968	pancreatic cancer	University of Texas Medical Branch at Galveston	mechanism of estrogen-induced carcinogenesis
Charles A. Janeway, Jr.	[1943-2003]	MD, 1969	B-cell lymphoma	Yale University	innate immunity and T lymphocyte biology
Edward Herbert	[1926-1987]	PhD, 1953	pancreatic cancer	Oregon Health & Science University	regulation of expression of opioid peptides and receptors
Thomas W. Smith	[1936-1997]	MD, 1965	mesothelioma	Harvard Medical School/Brigham & Women's Hospital	Mechanism and reversal studies of digitalis
Roy H. Steinberg	[1935-1997]	MD/PhD, 1965	multiple myeloma	University of California — San Francisco	pigment epithelium interactions with neural retina
David W. Fulker	[1937-1998]	PhD, 1967	pancreatic cancer	University of Colorado at Boulder	adoption studies of development in middle childhood
Donald J. Cohen	[1940-2001]	MD, 1966	ocular melanoma	Yale University	Tourette's syndrome and autism in children
Harvey D. Preisler	[1941-2002]	MD, 1965	lymphoma	Rush Medical College	clinical and biological studies of myeloid leukemias
Carl M. Pearson	[1919-1981]	MD, 1946	cancer	UCLA	studies in adjuvant-induced arthritis
Morton I. Grossman	[1919-1981]	MD/PhD, 1944	esophageal cancer	UCLA	studies on the etiology of peptic ulcer
Mones Berman	[1920-1982]	PhD, 1957	cancer	National Cancer Institute	quantitative, model-based problems in metabolism and endocrinology
Henry R. Mahler	[1921-1983]	PhD, 1948	heart failure	Indiana University	respiratory enzymes–structure, function, & biosynthesis
Milton Kern	[1925-1987]	PhD, 1954	lung cancer	NIH	ribonucleic acids of specifically isolated ribosomes
Thoralf M. Sundt, Jr.	[1930-1992]	MD, 1959	bone marrow cancer	Mayo Clinic	surgical techniques for intracranial aneurysms
John C. Liebeskind	[1935-1997]	PhD, 1962	cancer	UCLA	behavioral and electrophysiological studies of pain
Marian W. Fischman	[1939-2001]	PhD, 1972	colon cancer	Columbia University	behavioral pharmacology of cocaine
David S. Sigman	[1939-2001]	PhD, 1965	brain cancer	UCLA	enzymology and gene targeting
Charles D. Heidelberger	[1920-1983]	PhD, 1946	carcinoma of nasal sinus	University of Southern California Keck School of Medicine	effects of fluorinated pyrimidines on tumors
Sidney H. Ingbar	[1925-1988]	MD, 1947	lung cancer	Harvard Medical School/Beth Israel Medical Center	physiology of the thyroid gland and its clinical diseases
Kiichi Sagawa	[1926-1989]	MD/PhD, 1958	cancer	Johns Hopkins University School of Medicine	modelling the mechanics of cardiac chamber contraction
Sydney E. Salmon	[1936-1999]	MD, 1962	pancreatic cancer	University of Arizona	quantitative method for evaluating changes in myeloma tumor mass
Eva J. Neer	[1937-2000]	MD, 1963	breast cancer	Harvard Medical School/Brigham & Women's Hospital	regulation and cellular levels of G protein subunits
Lawrence D. Jacobs	[1938-2001]	MD, 1965	cancer	SUNY Buffalo	recombinant b interferon as treatment for Multiple Sclerosis
Richard J. Wyatt	[1939-2002]	MD, 1964	lung cancer	NIH	biochemistry of schizophrenia
Robert J. Fass	[1939-2002]	MD, 1964	lung cancer	Ohio State University	In vitro methods to test antimicrobial susceptibility of infectious agents
Michael Doudoroff	[1911-1975]	PhD, 1939	cancer	University of California — Berkeley	taxonomy and phylogeny of pseudomonads
Arnold M. Seligman	[1912-1976]	MD, 1937	prolonged terminal illness	Johns Hopkins University School of Medicine	drug development for prostatic carcinoma
Frederick H. Carpenter	[1918-1982]	PhD, 1944		University of California — Berkeley	mechanism of leucine aminopeptidase
Harvey M. Patt	[1918-1982]	PhD, 1942		University of California — San Francisco	ultra-high dose rates in experimental radiotherapy
Teruzo Konishi	[1920-1984]	MD/PhD, 1955	cancer	NIEHS	physiological and biophysical functions of the inner ear
Mortimer B. Lipsett	[1921-1985]	MD, 1951	brain tumor	NIH	steroid metabolic conversions in human subjects
Andrew C. Peacock	[1921-1985]	PhD, 1949	cancer	NIH/NCI	materials and methods for polyacrylamide gel electrophoresis
Harold Edelhoch	[1922-1986]	PhD, 1947	cancer	NIH/NIDDK	fluorescence methods for the study of protein structures
Gerald L. Klerman	[1928-1992]	MD, 1954	diabetes	Cornell University — Weill Medical College	phsychological studies of depression, schizophrenia and panic and other anxiety disorders
Nina S. Braunwald	[1928-1992]	MD, 1952	cancer	Harvard Medical School/Brigham & Women's Hospital	development of prosthetic heart valves for children
Amico Bignami	[1930-1994]	MD, 1954	brain cancer	Harvard Medical School	brain specific protein in astrocytes
Frank A. Oski	[1932-1996]	MD, 1958	prostate cancer	Johns Hopkins University School of Medicine	erythrocyte metabolism in the newborn infant
Richard P. Bunge	[1932-1996]	MD, 1960	esophageal cancer	University of Miami	schwann cell biology and human spinal cord injury
Harold C. Neu	[1934-1998]	MD, 1960	glioblastoma	Columbia University	surface enzymes in bacteria
Jiri Palek	[1934-1998]	MD, 1958	2 year illness	Tufts University	membrane properties of abnormal red cells
Irving Kupfermann	[1938-2002]	PhD, 1964	Creutzfeldt-Jacob's disease	Columbia University	Behavioral and neural analysis of learning in aplaysia
Merton Bernfield	[1938-2002]	MD, 1961	Parkinson's Disease	Harvard Medical School/Children's Hospital	nature and interactions of cell surface proteoglycans during morphogenesis
Eleanor M. Saffran	[1938-2002]	PhD, 1968	amyotrophic lateral sclerosis	Temple University School of Medicine	cognitive deficits in brain-damaged patients
Barbara J. Lowery	[1938-2002]	PhD, 1973	ovarian cancer	University of Pennsylvania School of Medicine	understanding stress responses of people who were physically ill
Elizabeth Stern	[1915-1980]	MD, 1940	cancer	UCLA	effects of steroid contraception on the ovary
Joseph Stokes, 3rd	[1924-1989]	MD, 1949	cancer	Boston University School of Medicine	epidemiological studies of coronary heart disease
W. Dean Warren	[1924-1989]	MD, 1950	cancer	Emory University	cirrhosis, shunt surgery, and nitrogen metabolism
Edward W. Purnell	[1928-1993]	MD, 1957	lung cancer	Case Western Reserve University School of Medicine	study of eye physiology and disease by ultrasound
Leo J. Neuringer	[1928-1993]	PhD, 1957	cancer	MIT	NMR studies of normal and transformed cell membranes
Frank Lilly	[1930-1995]	PhD, 1965	prostate cancer	Albert Einstein College of Medicine of Yeshiva University	role of hereditary factors in governing susceptibility to cancer-causing agents
Edwin L. Bierman	[1930-1995]	MD, 1955	bone cancer	University of Washington School of Medicine	Metabolism of particulate fat in diabetes and atherosclerosis
Kenneth W. Sell	[1931-1996]	MD/PhD, 1968	complications from diabetes	Emory University School of Medicine	human tissue banking and transplantation
Edgar Haber	[1932-1997]	MD, 1956	multiple myeloma	Harvard University School of Public Health	biological regulation of the renin-angiotensin system
J. Christian Gillin	[1938-2003]	MD, 1966	esophageal cancer	University of California — San Diego	serotenergic mechanisms in sleep and depression
Albert Dorfman	[1916-1982]	MD/PhD, 1944	kidney failure	University of Chicago	biochemistry of connective tissues
Henry S. Kaplan	[1918-1984]	MD, 1940	lung cancer	Stanford University	radiation-induced leukemia in the C57BL mouse
Charlotte Friend	[1921-1987]	PhD, 1950	lymphoma	Mount Sinai School of Medicine	tissue studies of murine virus-induced leukemia
William H. Tooley	[1925-1992]	MD, 1949	long illness	University of California — San Francisco	prevention and treatment of respiratory distress in neonates
Charles G. Moertel	[1927-1994]	MD, 1953	Hodgkin's Disease	Mayo Clinic	clinical treatments of gastrointestinal cancer
Barbara H. Bowman	[1930-1996]	PhD, 1959	cancer	University of Texas HSC at San Antonio	genetic control of the structure of human proteins
J. Calvin Giddings	[1930-1996]	PhD, 1955	prolonged battle with cancer	University of Utah	biomedical separations: field-flow fractionation

Investigator Name			Cause of death if known	Institution at the time of death	Scientific domain
John R. Williamson	[1934-2000]	PhD, 1959	cancer	University of Pennsylvania School of Medicine	molecular mechanisms of hormonal signal transduction
John S. O'Brien	[1934-2001]	MD, 1960	postpolio complications	University of California — San Diego	discovery of the gene responsible for Tay-Sachs disease
Jon I. Isenberg	[1937-2003]	MD, 1963	cancer	University of California — San Diego	duodenal mucosal bicarbonate secretion in human
George G. Glenner	[1927-1995]	MD, 1953	systemic senile amyloidosis	University of California — San Diego	molecular structure of the amyloid protein
J. Kiffin Penry	[1929-1996]	MD, 1955	complications of diabetes	Bowman Gray School of Medicine at Wake Forest University	controlled clinical trials of anticonvulsant and anti-epileptic drugs
Paul C. MacDonald	[1930-1997]	MD, 1955	cancer	University of Texas Southwestern Medical Center at Dallas	origin and interconversion of gonadal and adrenal streoid hormones
John Gibbon	[1934-2001]	PhD, 1967	cancer	Columbia University	CNS functions underlying the interval time sense in animals and humans
Donald F. Summers	[1934-2001]	MD, 1959	cancer	NIH	composition, assembly and replication of RNA viruses
R. Gordon Gould	[1910-1978]	PhD, 1933	cancer	Stanford University	internal medicine and cardiology
Sol Spiegelman	[1914-1983]	PhD, 1944	pancreatic cancer	Columbia University College of Physicians & Surgeons	nucleic acid hybridization
Frederick S. Philips	[1916-1984]	PhD, 1940	cancer	Sloan Kettering Institute for Cancer Research	pharmacological properties of chemotherapeutic agents and chemical carcinogenesis
Cyrus Levinthal	[1922-1990]	PhD, 1951	lung cancer	Columbia University College of Physicians & Surgeons	colinearity of genes and proteins, and the nature of messenger RNA
Sidney Leskowitz	[1923-1991]	PhD, 1950	brain tumor	Tufts University	cellular aspects of tolerance & delayed hypersensitivity
Kenneth M. Moser	[1929-1997]	MD, 1954	cancer	University of California — San Diego	clinical outcomes after pulmonary thromboendarterectomy
Donald A. Pious	[1930-1998]	MD, 1956	cancer	University of Washington School of Medicine	somatic cell genetic analysis of human immune response genes
Louis V. Avioli	[1931-1999]	MD, 1957	cancer	Washington University in St. Louis	mineral and skeletal metabolism in diabetes, kidney, and gastrointestinal disorders
Joseph E. Coleman	[1930-1999]	MD/PhD, 1963	cancer	Yale University	structure and function of metalloenzyme synthesis
Harvey C. Knowles, Jr.	[1915-1984]	MD, 1942	cancer	University of Cincinnati/Children's Hospital	clinical studies of gestational diabetes
Joseph Cochin	[1916-1985]	MD/PhD, 1955	leukemia	Boston University School of Medicine	factors in tolerance to the narcotic analgesics
Albert L. Lehninger	[1917-1986]	PhD, 1942	complications from asthma	Johns Hopkins University School of Medicine	structure and function of mitochondria
Charles W. Todd	[1918-1987]	PhD, 1943	long illness	City of Hope Medical Center	immunology & immunochemistry of tumor antigens
David H. Blankenhorn	[1924-1993]	MD, 1947	prostate cancer	University of Southern California Keck School of Medicine	control of risk factors in atherosclerosis
Paul M. Gallop	[1927-1996]	PhD, 1953	cancer	Harvard Medical School/Children's Hospital	Protein structure and collagen maturation
David J.L. Luck	[1929-1998]	MD/PhD, 1962	lymphoma	Rockefeller University	microtubular systems in human cells
Edward W. Moore	[1930-1999]	MD, 1955	aspergillosis	Medical College of Virginia	Pathophysiology of the billiary tract and gallbladder
Donald J. Reis	[1931-2000]	MD, 1956	hepatic cancer	Cornell University — Weill Medical College	neural control of blood circulation
Julius Marmur	[1926-1996]	PhD, 1951	lymphoma	Albert Einstein College of Medicine of Yeshiva University	genetics and biochemistry of cellular regulation
Nemat O. Borhani	[1926-1996]	MD, 1949	acute leukemia	University of Nevada at Reno	multicenter clinical studies of hypertension and cardiovascular disease
Russell Ross	[1929-1999]	DDS/PhD, 1962	cancer	University of Washington School of Medicine	response-to-injury origins of atherosclerosis
Richard A. Carleton	[1931-2001]	MD, 1955	cancer	Brown University Medical School	clinical studies of diet and smoking as cardiovascular disease risk factors
Gilda H. Loew	[1931-2001]	PhD, 1957	breast cancer	Molecular Research Institute	computational investigation of the structural and functional aspects of heme proteins and enzymes
N. Raphael Shulman	[1925-1996]	MD, 1947	cancer	NIH/NIDDK	mechanisms of autoimmune, alloimmune, and drug-dependent cytopenias
George Winokur	[1925-1996]	MD, 1947	pancreatic cancer	University of Iowa School of Medicine	genetics of bipolar disease, mania, alcoholism and other psychiatric diseases
Giovanni Di Chiro	[1926-1997]	MD, 1949	lung cancer	NIH	interventional neuroradiology
Norman P. Salzman	[1926-1997]	PhD, 1953	pancreatic cancer	NIH	glycosylation of SIV gp120–role in the immune response
Fritz E. Dreifuss	[1926-1997]	MD, 1950	lung cancer	University of Virginia School of Medicine	clinical investigations of childhood epilepsy
Dante G. Scarpelli	[1927-1998]	MD/PhD, 1960	esophageal adenocarcinoma	Northwestern University	metabolism of pancreatic carcinogens
Hans J. Müller-Eberhard	[1927-1998]	MD, 1953	cancer	Scripps Research Institute	identification of proteins and reaction mechanisms of the complement system
Miriam M. Salpeter	[1929-2000]	PhD, 1953	thyroid cancer	Cornell University	neurobiology of myasthenia gravis
Gerald Cohen	[1930-2001]	PhD, 1955	cancer	Mount Sinai School of Medicine	H2O2 and oxy-radical stress in catecholamine neurons
James K. McDougall	[1931-2003]	PhD, 1971	gastric cancer	University of Washington/FHCRC	role of DNA viruses in cancer
Edward H. Kass	[1917-1990]	MD/PhD, 1947	lung cancer	Harvard Medical School/Brigham & Women's Hospital	mechanism of toxic shock syndrome
Norman Kretchmer	[1923-1995]	MD/PhD, 1952	kidney cancer	University of California — Berkeley	regulation of metabolism during developement
Adolph I. Cohen	[1924-1996]	PhD, 1954	leukemia	Washington University in St. Louis	biochemistry and pharmacology of the retina
John L. Doppman	[1928-2000]	MD, 1953	cancer	NIH	flow dynamics in anterior spinal artery
David E. Green	[1910-1983]	PhD, 1934	cancer	University of Wisconsin	molecular biology of membrane systems
Alton Meister	[1922-1995]	MD, 1945	complications from a stroke	Cornell University — Weill Medical College	amino acid and glutathione biochemistry
Gisela Mosig	[1930-2003]	PhD, 1959	undergoing cancer treatment for two years	Vanderbilt University	dna replication and recombination in bacteriophages
Choh Hao Li	[1913-1987]	PhD, 1938	cancer of the pharynx	University of California — San Francisco	isolation and synthesis the human pituitary growth hormone
Robert H. Abeles	[1926-2000]	PhD, 1955	Parkinson's disease	Brandeis University	rational design of small-molecule inhibitors of enzymes
Alfred P. Wolf	[1923-1998]	PhD, 1953	lengthy illness	Brookhaven National Laboratory	synthesis of simple molecules in pure form and high specific activity for PET
Marian E. Koshland	[1921-1997]	PhD, 1949	lung cancer	University of California — Berkeley	biochemical methods to examine the immune response
Timothy J. Regan	[1924-2001]	MD, 1952	colon cancer	UMDNJ Newark	myocardial function and metabolism in chronic disease
Thomas C. Chalmers	[1917-1995]	MD, 1943	prostate cancer	Mount Sinai School of Medicine	inter-hospital cooperative studies of cirrhosis
Mortimer M. Elkind	[1922-2000]	PhD, 1953	long illness	Colorado State University	cell radiation response of cultured mammalian cells
Hamish N. Munro	[1915-1994]	MD/PhD, 1956	died in a nursing home. Parkinson	Tufts University	nutritional regulation of protein metabolism
Ruth Sager	[1916-1997]	PhD, 1948	bladder cancer	Harvard Medical School/DFCI	role of tumor suppressor genes in breast cancer
David M. Maurice	[1922-2002]	PhD, 1951	liver cancer	Columbia University College of Physicians & Surgeons	interference theory of corneal transparency
Robert A. Good	[1922-2003]	MD/PhD, 1947	esophageal cancer	University of South Florida College of Medicine	role of the thymus in immune system development
Harland G. Wood	[1907-1991]	PhD, 1935	lymphoma	Case Western Reserve University School of Medicine	heterotrophic carbon dioxide fixation
Hans Popper	[1903-1988]	MD/PhD, 1944	pancreatic cancer	Mount Sinai School of Medicine	correlation of structure and function in liver disease
Fritz A. Lipmann	[1899-1986]	MD/PhD, 1928	natural reasons	Rockefeller University	glucose transport in normal and malignant cells
Paul J. Scheuer	[1915-2003]	PhD, 1950	leukemia	University of Hawaii	structure and properties of spinochromes
Berta V. Scharrer	[1906-1995]	PhD, 1930	natural causes	Albert Einstein College of Medicine of Yeshiva University	immunocytochemical study of invertebrate nervous system
Michael W. Pozen	[1945-1981]	MD/PhD, 1974	heart attack	Boston University School of Medicine	confirmation parameters to assess EMT's decisions
Ronald E. Talcott	[1947-1984]	PhD, 1973	automobile accident	University of California — San Francisco	carboxylesterases of toxicologic significance
Nathaniel A. Young	[1939-1979]	MD, 1962	drowned in British Virgin Islands	National Cancer Institute	oncology and molecular pathology
Ahmad I. Bukhari	[1943-1983]	PhD, 1971	heart attack	Cold Spring Harbor Laboratory	life cycle of mutator phage μ
Alan P. Wolffe	[1959-2001]	PhD, 1984	car accident	NIH	role of DNA methylation in regulating gene expression in normal and pathological states
Shu-Ren Lin	[1936-1979]	MD, 1962	plane crash	University of Rochester	imaging studies of cerebral blood flow after cardiac arrest
William D. Nunn	[1943-1986]	PhD, 1972	sudden cardiac arrest	University of California — Irvine	regulation of fatty acid/acetate metabolism in e. coli
John L. Kemink	[1949-1992]	MD, 1975	murder	University of Michigan, Ann Arbor	vestibular diagnosis and surgery, acoustic neuromas, and cochlear implants
Stanley R. Kay	[1946-1990]	PhD, 1980	heart attack	Albert Einstein College of Medicine of Yeshiva University	symptoms and diagnostic tests of schizophrenia
Roberta D. Shahin	[1953-1997]	PhD, 1985	sudden accute illness	Center for Biologics Evaluation and Research	mouse model of respiratory B. pertussis infection in mice
Robert M. Pratt, Jr.	[1942-1987]	PhD, 1970	died in his sleep	NIEHS/University of North Carolina at Chapel Hill	craniofacial development of the fetus
Howard J. Eisen	[1942-1987]	MD, 1969	suicide	NIH/NICHD	mechanism of action of cortisol and related glucocorticoid hormones
Joaquim Puig-Antich	[1944-1989]	MD, 1967	asthma attack	University of Pittsburgh	psychobiology and treatment of child depression
Elizabeth A. Rich	[1952-1998]	MD, 1977	traffic accident	Case Western Reserve University School of Medicine	natural history of lymphocytic alveolitis in hiv disease
Jeffrey M. Hoeg	[1952-1998]	MD, 1977	renal cancer	NIH/NHLBI	lipoprotein metabolism and its connection to cardiovascular disease
Matthew L. Thomas	[1953-1999]	PhD, 1981	died while travelling	Washington University in St. Louis	function and regulation of leukocyte surface glycoproteins
Mu-En Lee	[1954-2000]	MD/PhD, 1984	complications from routine surgery	Harvard Medical School/MGH	characterization of vascular smooth muscle LIM protein
Tsunao Saitoh	[1949-1996]	PhD, 1977	murdered	University of California — San Diego	altered protein kinases in alzheimer's disease
James W. Prahl	[1931-1979]	MD/PhD, 1964	rock climing accident	University of Utah	structural basis of the functions of human complement
Pokar M. Kabra	[1942-1990]	PhD, 1972	plane crash	University of California — San Francisco	application of liquid chromatography to therapeutic drug monitoring
Harold A. Menkes	[1938-1987]	MD, 1963	car accident	Johns Hopkins University School of Medicine	occupational and environmental lung disease

Investigator Name			Cause of death if known	Institution at the time of death	Scientific domain
Richard E. Heikkila	[1942-1991]	PhD, 1969	murder	UMDNJ Robert Wood Johnson Medical School	oxidation-reduction reactions and the dopamine receptor system
Howard S. Tager	[1945-1994]	PhD, 1971	heart attack	University of Chicago	biochemical structure, action, regulation and degradation of the insulin and glucagon molecules
Sukdeb Mukherjee	[1946-1995]	MD, 1971	short illness	Medical College of Georgia	neuroleptic effects on regional cerebral blood flow
John J. Wasmuth	[1946-1995]	PhD, 1973	heart attack	University of California — Irvine	human-hamster somatic cell hybrids/localization of Hnyington's disease gene
Richard P. Nordan	[1949-1998]	PhD, 1983	cerebral aneurysm	NIH	immunologist and molecular biologist
Roland L. Phillips	[1937-1987]	MD/PhD, 1971	glider plane accident	Loma Linda University School of Medicine	role of lifestyle in cancer and cardiovascular disease among Adventists
Samuel A. Latt	[1938-1988]	MD/PhD, 1971	heart attack	Harvard Medical School/Children's Hospital	genetic and cytogenetic studies of mental retardation
Emil T. Kaiser	[1938-1988]	PhD, 1959	complications from kidney transplant	Rockefeller University	mechanism of carboxypeptidase action
D. Michael Gill	[1940-1990]	PhD, 1967	heart attack	Tufts University	biochemistry of cholera toxin and other pathogenic toxins
John P. Merlie	[1945-1995]	PhD, 1973	heart failure	Washington University in St. Louis	molecular genetics of the acetylcholine receptor
Robert S. Krooth	[1929-1980]	MD/PhD, 1957	suicide/self-inflicted gunshot wound	Columbia University College of Physicians & Surgeons	biochemical deffects in inherited metabolic disorders
Takeo Kakunaga	[1937-1988]	PhD, 1966	lung cancer with a brain metastasis	NIH/NCI	malignant transformation of mammalian cells by chemical carcinogens
Abraham Worcel	[1938-1989]	MD, 1963	suicide	University of Rochester	structure of interphase and metaphase chromosomes
Roland D. Ciaranello	[1943-1994]	MD, 1970	heart attack	Stanford University	molecular neurobiology and developmental disorders
Gary J. Miller	[1950-2001]	MD/PhD, 1978	heart attack	University of Colorado Health Sciences Center	vitamin D receptors in the growth regulation of prostate cancer cells
William B. Reed	[1924-1976]	MD, 1952		University of Southern California Keck School of Medicine	cutaneous genetic disorders
James R. Neely	[1936-1988]	PhD, 1966	heart attack	Penn State University	effects of diabetes and oxygen deficiency in regulation of metabolism in the heart
Mary Lou Clements	[1946-1998]	MD, 1972	airplane crash	Johns Hopkins University School of Medicine	development of AIDS vaccines
John B. Penney, Jr.	[1947-1999]	MD, 1973	heart attack	Harvard Medical School/MGH	receptor mechanisms in movement disorder pathophysiology
Lynn M. Wiley	[1947-1999]	PhD, 1975	plane crash	University of California — Davis	morphogenesis in early mammalian embryos
Trudy L. Bush	[1949-2001]	PhD, 1977	heart attack	University of Maryland School of Medicine	postmenopausal estrogen/progestins interventions
Arend Bouhuys	[1926-1979]	MD/PhD, 1956	heart attack	Yale University	community studies of obstructive lung disease
Erhard Gross	[1928-1981]	PhD, 1958	automobile collision	NIH/NICHD	structural analysis of naturally-occuring peptide antibiotics
Richard C. Lillehei	[1928-1981]	MD/PhD, 1960	died while jogging	University of Minnesota	mechanisms of RES stimulation in experimental shock
Hymie L. Nossel	[1930-1983]	MD/PhD, 1962	heart attack	Columbia University	causes of thrombosis and the nature of hemostasis
James C. Steigerwald	[1935-1988]	MD, 1961		University of Colorado Health Sciences Center	internal medicine / rheumatology
Simon J. Pilkis	[1942-1995]	MD/PhD, 1971	heart attack	University of Minnesota	carbohydrate metabolism and diabetes
James Olds	[1922-1976]	PhD, 1952	swimming accident	California Institute of Technology	pharmacology of motivational mechanisms
Peter W. Neurath	[1923-1977]	PhD, 1950	heart attack	Tufts University	chromosomal variants of cells converted by viruses
Emanuel M. Bogdanove	[1925-1979]	PhD, 1953	killed in an accident	Medical College of Virginia	endocrine-influencing centers in the hypothalamus
Harold A. Baltaxe	[1931-1985]	MD, 1960	heart attack	University of California — Davis	development of new coronary angiographic techniques
Roy D. Schmickel	[1936-1990]	MD, 1961	died tragically	University of Pennsylvania School of Medicine	isolation and characterization of human ribosomal DNA
Fredric S. Fay	[1943-1997]	PhD, 1969	heart attack	UMASS	generation and regulation of force in smooth muscle
Roger R. Williams	[1944-1998]	MD, 1971	airplane crash	University of Utah	genetics and epidemiology of coronary artery diseases
Jeffrey M. Isner	[1947-2001]	MD, 1973	heart attack	Tufts University	therapeutic angiogenesis in vascular medicine, cardiovascular laser phototherapy
Gustavo Cudkowicz	[1927-1982]	MD, 1952	brief illness	SUNY Buffalo	controls of proliferation specific for leukemias
John C. Seidel	[1933-1988]	PhD, 1961	heart attack	Boston Biomedical Research Institute	actin-myosin interaction in pulmonary smooth muscle
William L. McGuire	[1937-1992]	MD, 1964	scuba-diving accident	University of Texas HSC at San Antonio	mechanisms of hormonal control and growth and regression of mammary carcinoma
Eric Holtzman	[1939-1994]	PhD, 1964	ingestion of potassium cyanide, self-administered	Columbia University	dynamic of cell membranes
Julio V. Santiago	[1942-1997]	MD, 1967	heart attack	Washington University in St. Louis	role of social factors, lifestyle practices, and medication in the onset of type II diabetes
John J. Pisano	[1929-1985]	PhD, 1955	heart attack	NIH/NHLBI	isolation of active peptides
Dale E. McFarlin	[1936-1992]	MD, 1961	heart attack	NIH	neuroimmunological studies of multiple sclerosis
Walter F. Heiligenberg	[1938-1994]	PhD, 1964	plane crash	University of California — San Diego	neuroethological studies of electrolocation
George J. Schroepfer, Jr.	[1932-1998]	MD/PhD, 1961	heart attack	Rice University	regulation of the formation and metabolism of cholesterol
Thomas A. McMahon	[1943-1999]	PhD, 1970	complications from routine surgery	Harvard University	orthopedic biomechanics
Joseph F. Foster	[1918-1975]	PhD, 1943	heart attack	Purdue University	configurational changes in protein molecules
Gerald P. Rodnan	[1927-1983]	MD, 1949	complications after vascular surgery	University of Pittsburgh	renal transport if uric acid and protein
George Streisinger	[1927-1984]	PhD, 1953	scuba-diving accident	University of Oregon	genetic mutations and the nervous system development in lower vertebrates
Lucien B. Guze	[1928-1985]	MD, 1951	sudden cardiac arrest	UCLA	pathogenesis of experimental pyelonephritis
Lubomir S. Hnilica	[1929-1986]	PhD, 1952	automobile accident	Vanderbilt University	nuclear antigens in human colorectal cancer
Charles L. Wittenberger	[1930-1987]	PhD, 1959	motorcycle accident	NIH/NINDR	regulation of the pathways of intermediary metabolism
D. Martin Carter	[1936-1993]	MD/PhD, 1971	dissecting aortic aneurysm	Rockefeller University	susceptibility of pigment and cutaneous cells to DNA injury by UV
Verne M. Chapman	[1938-1995]	PhD, 1965	died suddenly while attending meeting	Roswell Park Cancer Institute/SUNY Buffalo	development of cumulative multilocus map of mouse chromosomes
Dolph O. Adams	[1939-1996]	MD/PhD, 1969	unexpected	Duke University	development and regulation of macrophage activation
Lee A. Lillard	[1943-2000]	PhD, 1972	heart attack	University of Michigan, Ann Arbor	aging and retirement studies
Don C. Wiley	[1944-2001]	PhD, 1971	accidental fall	Harvard University	viral membrane and glycoprotein structure
Lonnie D. Russell, Jr.	[1944-2001]	PhD, 1974	swimming accident	Southern Illinois University School of Medicine	filament regulation of spermatogenesis
Herbert J. Rapp	[1923-1981]	PhD, 1955		National Cancer Institute	immunologist and cancer research
Eugene C. Jorgensen	[1923-1981]	PhD, 1953	murdered	University of California — San Francisco	structure/activity relationships of compounds related to thyroxin
Margaret O. Dayhoff	[1925-1983]	PhD, 1948	heart attack	Georgetown University Medical Center	computer study of sequences of amino acids in proteins
Norman Geschwind	[1926-1984]	MD, 1951	heart attack	Harvard Medical School/Beth Israel Medical Center	relationship between the anatomy of the brain and behavior
Laurence M. Sandler	[1929-1987]	PhD, 1956	heart attack	University of Washington School of Medicine	cytogenetics of meiosis and development in drosophila
L. Rao Chervu	[1930-1988]	PhD, 1962	brutally murdered	Albert Einstein College of Medicine of Yeshiva University	improved radiopharmaceuticals for nephrology and urology
Peter M. Steinert	[1945-2003]	PhD, 1972	heart attack	NIH	structures and interactions of the proteins characteristic of epithelial cells
Arnold Lazarow	[1916-1975]	MD/PhD, 1941	brief illness	University of Minnesota	fetal endocrinology and study of diabetes & pregnancy
Edward V. Evarts	[1926-1985]	MD, 1948	heart attack	NIH	electrophysiological activity of in vivo neurons in waking and sleeping states
Anthony Dipple	[1940-1999]	PhD, 1964	heart attack	NIH	metabolic activation and DNA interactions of polycyclic aromatic hydrocarbon carcinogens
Gerald L. Stoner	[1943-2002]	PhD, 1974	complications following a fall	NIH/NINDS	neuropathology and molecular epidemiology of the human polyomavirus
G. Scott Giebink	[1944-2003]	MD, 1969	heart attack	University of Minnesota	pathogenesis of otitis media and immunizations
Daniel A. Brody	[1915-1975]	MD, 1940	heart attack	University of Tennessee	generator properties of isolated mammalian hearts
Michelangelo G.F. Fuortes	[1917-1977]	MD, 1941		NIH/NINDS	study of the peripheral visual system in vertebrate animals
Sidney Riegelman	[1921-1981]	PhD, 1948	drowned while scuba diving	University of California — San Francisco	intersubject variation in first pass effect of drugs
Lewis W. Wannamaker	[1923-1983]	MD, 1948	heart attack	University of Mississippi Medical Center	clinical and epidemiologic aspects of streptococcal infections
Donald J. Magilligan, Jr.	[1929-1989]	MD, 1965	short illness	Henry Ford Health Sciences Center	natural history and limitations of porcine heart valves
Ronald G. Thurman	[1941-2001]	PhD, 1967	massive heart attack	University of North Carolina at Chapel Hill	hepatic metabolism, alcoholic liver injury and toxicology
F. Brantley Scott, Jr.	[1930-1991]	MD, 1955	plane crash	Baylor University College of Medicine/St. Luke's Episcopal Hospital	development of the penile prosthesis
DeWitt S. Goodman	[1930-1991]	MD, 1955	pulmonary embolism	Columbia University	lipid metabolism and its role in the development of heart and artery disease
Donald C. Shreffler	[1933-1994]	PhD, 1961	heart attack	Washington University in St. Louis	organization and functions of H-2 gene complex
A. Arthur Gottlieb	[1937-1998]	MD, 1961	pulmonary embolus following surgery	Tulane University School of Medicine	role of macrophage nucleic acid in antibody production
John N. Whitaker	[1940-2001]	MD, 1965	injuries following a bycicle race	University of Alabama at Birmingham	molecular immunopathogenesis of demyelinating disease
Christopher A. Dawson	[1942-2003]	PhD, 1969	suddenly	Medical College of Wisconsin	pulmonary hemodynamics
Maurice S. Raben	[1915-1977]	MD, 1939		Tufts University	humoral and metabolic aspects of cardiac function
Josiah Brown	[1923-1985]	MD, 1947	tragic accident	UCLA	biochemical studies of lipid and carbohydrate metabolism
John H. Walsh	[1938-2000]	MD, 1963	heart attack	UCLA	gastrointestinal hormones, gastric acid production and peptic ulcer disease
Jerome R. Vinograd	[1913-1976]	PhD, 1940		California Institute of Technology	biochemistry and molecular biology

Investigator Name			Cause of death if known	Institution at the time of death	Scientific domain
Merton F. Utter	[1917-1980]	PhD, 1942		Case Western Reserve University School of Medicine	structure and function of pep carboxykinase isozymes
E. Jack Wylie	[1918-1982]	MD, 1943	heart attack	University of California — San Francisco	development of techniques for the treatment and management of chronic visceral ischemia
Kwan C. Tsou	[1922-1985]	PhD, 1950	heart attack	University of Pennsylvania School of Medicine	development of serum nuclease isozyme test for cancer
Norbert Freinkel	[1926-1989]	MD, 1949	heart attack	Northwestern University	metabolic regulation in normal and diabetic pregnancies
Edgar C. Henshaw	[1929-1992]	MD, 1956	complications from early-stage cancer treatment	University of Rochester	intermediary metabolism in animals and in man
Donald T. Witiak	[1935-1998]	PhD, 1961	stroke	University of Wisconsin	stereochemical studies of hypocholesterolemic agents
Thomas P. Dousa	[1937-2000]	MD/PhD, 1968	heart attack	Mayo Clinic	cellular action of vasopressin in the kidney
Thomas F. Burks, II	[1938-2001]	PhD, 1967	heart attack	University of Texas HSC at Houston	central and peripheral neuropeptide pharmacology
Robert M. Macnab	[1940-2003]	PhD, 1969	accidental fall	Yale University	sequence analysis and function of bacterial flagellar motor
David Pressman	[1916-1980]	PhD, 1940		Roswell Park Cancer Institute/SUNY Buffalo	structure and function of antibody molecules and tissue antigens of the HLA system
Abraham M. Lilienfeld	[1920-1984]	MD, 1944	heart attack	Johns Hopkins University School of Public Health	epidemiological methods for the study of chronic diseases
Marion I. Barnhart	[1921-1985]	PhD, 1950	traffic accident	Wayne State University School of Medicine	cellular sites for synthesis of blood proteins
Thomas R. Johns, 2nd	[1924-1988]	MD, 1948	refractory arrhythmia	University of Virginia School of Medicine	physiological studies of myasthenia gravis
Gerald D. Aurbach	[1927-1991]	MD, 1954	hit in a head by a stone	NIH	bone metabolism and calcium homeostasis
Demetrios Papahadjopoulos	[1934-1998]	PhD, 1963	adverse drug reaction/multi-organ failure	University of California — San Francisco	phospholipid-protein interactions, lipid vesicles, and membrane function
Takis S. Papas	[1935-1999]	PhD, 1970	unexpected and sudden	Medical University of South Carolina	characterization of ETS genes and retroviral onc genes
John J. Jeffrey, Jr.	[1937-2001]	PhD, 1965	stroke	Albany Medical College	mechanism of action and the physiologic regulation of mammalian collagenases
Victor J. Ferrans	[1937-2001]	MD/PhD, 1963	complications from diabetes	NIH	myocardial and vascular pathobiology
James N. Davis	[1939-2003]	MD, 1965	airplane crash	SUNY HSC at Stony Brook	mechanisms underlying neuronal injury after brain ischemia
Frederick B. Bang	[1916-1981]	MD, 1939	heart attack	Johns Hopkins University School of Medicine	cell virus relationships in respiratory mucosae
James M. Felts	[1923-1988]	PhD, 1955	heart failure	University of California — San Francisco	synthesis and processing of plasma lipoproteins
Ernst Freese	[1925-1990]	PhD, 1954	cerebral hemorrhage	NIH/NINDS	studies of environmental mutagenesis
Lucien J. Rubinstein	[1924-1990]	MD, 1948	ruptured intracranial aneurysm	University of Virginia School of Medicine	differentiation and stroma-induction in neural tumors
George B. Craig, Jr.	[1930-1995]	PhD, 1956	heart attack	University of Notre Dame	genetics and reproductive biology of aedes mosquitoes
James R. Klinenberg	[1934-1999]	MD, 1959	intracerebral hemorrhage	UCLA	pathophysiology of gout and hyperuricemia
Paul B. Sigler	[1934-2000]	MD/PhD, 1967	heart attack	Yale University	structural analysis of biological macromolecules
Sandy C. Marks, Jr.	[1937-2002]	DDS/PhD, 1968	heart attack	UMASS	vitamin D and bone modeling
Albert H. Coons	[1912-1978]	MD, 1937	coronary disease and congestive heart failure	Harvard Medical School	studies on antibody formation
Henry G. Kunkel	[1916-1983]	MD, 1942	complications after vascular surgery	Rockefeller University	identification of MHC Class II molecules
Edgar E. Ribi	[1920-1986]	PhD, 1948	plane crash	NIH/NIAID	fine structure of immunologically-active cell constituents for the development of vaccines
Bertram Sacktor	[1922-1988]	PhD, 1949	heart attack	National Institute on Aging in Baltimore	mechanisms of hormonal regulation of cellular pH and mineral metabolism in the kidney
Lucille S. Hurley	[1922-1988]	PhD, 1950	complications from open heart surgery	University of California — Davis	genetic and nutritional interactions in development
Paul Margolin	[1923-1989]	PhD, 1956	heart attack	City College of New York	mutation and suppressor studies of a bacterial gene
Zanvil A. Cohn	[1926-1993]	MD, 1953	aortic dissection	Rockefeller University	macrophage in cell biology and resistance to infectious disease
Carl Monder	[1928-1995]	PhD, 1956	brief illness, acute fulminating leukemia	Population Council	corticosteroid metabolism in juvenile hypertension
Gordon Guroff	[1933-1999]	PhD, 1959	car accident	NIH/NICHD	biochemical and molecular biological studies of nerve growth factor
Gerald P. Murphy	[1934-2000]	MD, 1959	heart attack	Roswell Park Cancer Institute/SUNY Buffalo	detection, immunotherapy, and prognostic indicators of prostate cancer
Alvito P. Alvares	[1935-2001]	PhD, 1966	killed by a car	Uniformed Services University of the Health Sciences	biochemical manifestations of toxicity in gold therapy
Patricia S. Goldman-Rakic	[1937-2003]	PhD, 1963	struck by a car	Yale University	development and plasticity of the primate frontal lobe
Stephen W. Kuffler	[1913-1980]	MD, 1937	heart attack	Harvard University	microphysiology of synaptic transmission
John P. Merrill	[1917-1984]	MD, 1942	drowned	Harvard Medical School/Brigham & Women's Hospital	role of the immune system in kidney transplantation
Abraham I. Braude	[1917-1984]	MD/PhD, 1950	heart attack	University of California — San Diego	pathogenesis and treatment of life-threatening septic shock
Susumu Hagiwara	[1922-1989]	PhD, 1951	bacterial infection	UCLA	evolutionary and developmental properties of calcium channels in cell membranes
Daniel Rudman	[1927-1994]	MD, 1949	complications from brain surgery	Medical College of Wisconsin	adipokinetic substances of the pituitary gland
Thomas G. Smith, Jr.	[1931-1998]	MD, 1960	heart attack	NIH/NINDS	fractal analysis of central nervous system neuron and glial cell morphology
Richard N. Lolley	[1933-2000]	PhD, 1961	heart attack	University of Southern California Keck School of Medicine	maturation of metabolism in normal & dystrophic retina
Joseph H. Ogura	[1915-1983]	MD, 1941	heart attack	Washington University in St. Louis	physiology of the larynx analog
Manfred M. Mayer	[1916-1984]	PhD, 1946	heart attack	Johns Hopkins University School of Medicine	immunochemistry of the complement system
Albert Segaloff	[1917-1985]	MD, 1942	brief illness	Tulane University School of Medicine	hormonal treatment of advanced breast cancer
F. Blair Simmons	[1930-1998]	MD, 1956	heart attack	Stanford University	development of a cochlear prothesis system for hearing loss
Henryk M. Wisniewski	[1931-1999]	MD/PhD, 1960	heart failure	SUNY Downstate Medical Center College of Medicine	pathogenesis of inflammatory demyelinating diseases
V. Everett Kinsey	[1909-1978]	PhD, 1937	stroke	Institute of Biological Sciences at Oakland University	intraocular fluid dynamics
Frederic C. Bartter	[1914-1983]	MD, 1940	stroke	University of Texas HSC at San Antonio	interaction between the kidney and various endocrine systems
Nathan O. Kaplan	[1917-1986]	PhD, 1943		University of California — San Diego	isolation and structure determination of coenzyme A
David T. Imagawa	[1922-1991]	PhD, 1950	heart attack	Harbor-UCLA Medical Center	morphological conversion with leukemia viruses
Robert H. Williams	[1909-1979]	MD, 1934	on an airline en route to Philadelphia	University of Washington School of Medicine	diabetes etiology, pathogenesis, and management
Toichiro Kuwabara	[1920-1991]	MD/PhD, 1952	heart failure	Harvard Medical School	ultrastructure of retina and retinal disease
William F. Harrington	[1920-1992]	PhD, 1952	heart failure	Johns Hopkins University School of Medicine	myosin thick filament structure and assembly
G. Jeanette Thorbecke	[1929-2001]	MD/PhD, 1954	stung by a Portuguese man-of-war jellyfish	New York University School of Medicine	histologic and functional aspects of lymphoid tissue development
Felix T. Rapaport	[1929-2001]	MD, 1954	coronary heart disease	SUNY HSC at Stony Brook	induction of unresponsiveness to allografts
Marian W. Kies	[1915-1988]	PhD, 1944	pancreatitis	NIH/MIMH	study of experimental allergic encephalomyelitis
Menek Goldstein	[1924-1997]	PhD, 1955	stroke	New York University School of Medicine	purification of enzymes in the catecholamine synthetic pathway
Andrew P. Somlyo	[1930-2003]	MD, 1956	heart attack	University of Virginia School of Medicine	vasomotor function of smooth muscle and their relation to heart disease
Koloman Laki	[1909-1983]	PhD, 1936	heart attack	NIH/NIDDK	purification of fibrinogen
Paul A. Srere	[1925-1999]	PhD, 1951	complications from liver surgery	University of Texas Southwestern Medical Center at Dallas	cell metabolism and the krebs tca cycle
D. Eugene Strandness, Jr.	[1928-2002]	MD, 1954	pulmonary failure	University of Washington School of Medicine	ultrasonic duplex scanner for noninvasive vascular disease diagnosis
Vincent Massey	[1926-2002]	PhD, 1953	heart attack	University of Michigan, Ann Arbor	biological oxidation mechanisms of proteins that contain riboflavin
Murray B. Bornstein	[1918-1995]	MD, 1952	cardiac aneurysm	Albert Einstein College of Medicine of Yeshiva University	copolymer as a protective treatment for the exacerbation of multiple sclerosis
Clarence J. Gibbs, Jr.	[1924-2001]	PhD, 1962	cardiac disease	NIH/NINDS	infectuous diseases of the nervous system
Russell L. De Valois	[1926-2003]	PhD, 1952	automobile accident	University of California — Berkeley	brain mechanisms underlying color vision
Efraim Racker	[1913-1991]	MD, 1938	stroke	Cornell University	identifying and purifying Factor 1, the first part of the ATP synthase enzyme
Walsh McDermott	[1901-1981]	MD, 1934	heart attack	Cornell University Medical College	latent and dormant microbial infections
Jonas E. Salk	[1914-1995]	MD, 1939	heart failure	Salk Institute	effective vaccine for polio
Lawrence Bogorad	[1921-2003]	PhD, 1949	stroke while on vacation	Harvard University	determinants of transcript longevity
Herman M. Kalckar	[1908-1991]	MD/PhD, 1939	pneumonia	Boston University School of Medicine	genes, enzymes, nucleotides, and carbohydrate patterns
Eugene M. Farber	[1917-2000]	MD, 1943	brief illness	Stanford University	biologic effects of photochemotherapy in psoriasis
Henry Rapoport	[1918-2002]	PhD, 1943	pneumonia	University of California — Berkeley	total synthesis of heterocyclic drugs
Norman R. Davidson	[1916-2002]	PhD, 1939	brief illness	California Institute of Technology	physical chemistry of nucleic acids
Karl A. Folkers	[1906-1997]	PhD, 1931	heart failure	University of Texas at Austin	peptide antagonists of LHRH as gonadotropin inhibitors
Margaret J. Sullivan	[1957-2001]	PhD, 1986		University of Missouri at Columbia	role of peptide neurotransmitters in body fluid homeostasis
Leonard R. Axelrod	[1927-1975]	PhD, 1952		Environmental Protection Agency	studies in steroid intermediate metabolism
Sidney R. Cooperband	[1931-1979]	MD, 1956		Boston University School of Medicine	lymphocyte proliferation inhibitory factor
James L. Lehr	[1940-1989]	MD, 1968		University of Chicago	modular computer-mediated radiology system
Alberto DiMascio	[1928-1978]	PhD, 1966		Tufts University	follow-up of maintenance treatment for depression
William B. Kinter	[1926-1978]	PhD, 1955		Mount Desert Island Biological Lab	membrane toxicity theory and environmental pollutants

Investigator Name			Cause of death if known	Institution at the time of death	Scientific domain
Alfred A. Smith	[1928-1980]	MD, 1956		New York Medical College	respiratory-depressive effects of ethanol
Leah M. Lowenstein	[1931-1984]	MD/PhD, 1958		Thomas Jefferson University Medical College	regulation of renal compensatory adaptation
S. Morris Kupchan	[1922-1976]	PhD, 1945		University of Virginia School of Medicine	chemistry of tumor-inhibitory natural products
Edward C. Heath	[1930-1985]	PhD, 1955		University of Iowa School of Medicine	molecular biology of tumor cells
Arnold F. Brodie	[1923-1981]	PhD, 1952		University of Southern California Keck School of Medicine	mechanisms of oxidative energy generation in bacteria
Alvin Nason	[1919-1978]	PhD, 1952		Johns Hopkins University School of Medicine	enzymology of nitrate respiration and assimilation
Andrew G. Morrow	[1923-1982]	MD, 1946		NIH/NHLBI	surgical correction of obstructive subaortic hypertrophy
Elijah Adams	[1918-1979]	MD, 1942		University of Maryland School of Medicine	tyrosinases and tyrosine hydroxylases
Myron L. Bender	[1924-1988]	PhD, 1948		Northwestern University	mechanism of action of proteases
Kenneth J.W. Taylor	[1939-2003]	MD/PhD, 1975		Yale University	diagnostic ultrasound imaging
Brigitte A. Prusoff	[1926-1991]	PhD, 1978		Yale University	follow-up of maintenance treatment for depression
Edwin D. Murphy	[1917-1984]	MD, 1943		NIH/NCI	gene mechanisms in autoimmunity and lymphoproliferation
Henry Kamin	[1920-1988]	PhD, 1948		Duke University	biological oxidations in mitochondria and microsomes
Henry A. Schroeder	[1906-1975]	MD, 1933		Dartmouth Medical School	abnormal trace metals in cardiovascular diseases
Carl L. Larson	[1909-1978]	MD, 1939		University of Montana at Missoula	specific and nonspecific resistance caused by t. bacilli
David F. Waugh	[1915-1984]	PhD, 1940		MIT	protein interactions and physico-chemical properties
John W. Porter	[1915-1984]	PhD, 1942		University of Wisconsin	regulation of lipogenesis by insulin and glucagon
Thomas F. Gallagher	[1905-1975]	PhD, 1931		Albert Einstein College of Medicine of Yeshiva University	metabolic transformation of steroid hormones
Benjamin Alexander	[1908-1978]	MD, 1934		NY Blood Center	coagulation, hemorrhage, and thrombosis
Bernard Saltzberg	[1919-1989]	PhD, 1972		University of Houston	electrophysiological analysis of learning disabilities
Georges Ungar	[1906-1977]	MD, 1939		University of Tennessee	chemical transfer of drug tolerance and learned behavior
Harold Koenig	[1921-1992]	MD/PhD, 1949		Northwestern University	molecular mechanisms of blood-brain barrier dysfunction
Albert S. Kaplan	[1917-1989]	PhD, 1952		Vanderbilt University	metabolism of cells infected with nuclear DNA viruses
Tsoo E. King	[1917-1990]	PhD, 1949		University of Pennsylvania School of Medicine	bioenergetic apparatus in heart mitochondria
Arthur Cherkin	[1913-1987]	PhD, 1953		Sepulveda VA Medical Center	role of cholinergic drugs in reducing the memory loss
Peter D. Klein	[1927-2001]	PhD, 1954		Baylor College of Medicine	metabolism of 13C compounds in digestive diseases
Alex B. Novikoff	[1913-1987]	PhD, 1938		Albert Einstein College of Medicine of Yeshiva University	histochemical studies of the Golgi apparatus
Walter E. Brown	[1918-1993]	PhD, 1949		American Dental Association Health Foundation	chemistry of calcium phosphates
C. Clark Cockerham	[1921-1996]	PhD, 1952		North Carolina State University	the statistics of genetic systems
Leo T. Samuels	[1899-1978]	PhD, 1930		University of Utah	steroid hormone metabolism and tumorogenic action
Peter N. Magee	[1921-2000]	MD, 1945		Thomas Jefferson University Medical College	genetic basis of carconogenesis

Appendix B: Linking Scientists with their Journal Articles

The source of our publication data is *PubMed*, a bibliographic database maintained by the U.S. National Library of Medicine that is searchable on the web at no cost.[iv] *PubMed* contains over 14 million citations from 4,800 journals published in the United States and more than 70 other countries from 1950 to the present. The subject scope of this database is biomedicine and health, broadly defined to encompass those areas of the life sciences, behavioral sciences, chemical sciences, and bioengineering that inform research in health-related fields. In order to effectively mine this publicly-available data source, we designed PubHarvester, an open-source software tool that automates the process of gathering publication information for individual life scientists (see Azoulay et al. 2006 for a complete description of the software). PubHarvester is fast, simple to use, and reliable. Its output consists of a series of reports that can be easily imported by statistical software packages.

This software tool does not obviate the two challenges faced by empirical researchers when attempting to accurately link individual scientists with their published output. The first relates to what one might term "Type I Error," whereby we mistakenly attribute to a scientist a journal article actually authored by a namesake; The second relates to "Type II error," whereby we conservatively exclude from a scientist's publication roster legitimate articles:

Namesakes and popular names. *PubMed* does not assign unique identifiers to the authors of the publications they index. They identify authors simply by their last name, up to two initials, and an optional suffix. This makes it difficult to unambiguously assign publication output to individual scientists, especially when their last name is relatively common.

Inconsistent publication names. The opposite danger, that of recording too few publications, also looms large, since scientists are often inconsistent in the choice of names they choose to publish under. By far the most common source of error is the haphazard use of a middle initial. Other errors stem from inconsistent use of suffixes (Jr., Sr., 2nd, etc.), or from multiple patronyms due to changes in spousal status.

To deal with these serious measurement problems, we opted for a labor-intensive approach: the design of individual search queries that relies on relevant scientific keywords, the names of frequent collaborators, journal names, as well as institutional affiliations. We are aided in the time-consuming process of query design by the availability of a reliable archival data source, namely, these scientists' CVs and biosketches. PubHarvester provides the option to use such custom queries in lieu of a completely generic query (e.g, `"azoulay p"[au]` or `"graff zivin js"[au]`). As an example, one can examine the publications of Scott A. Waldman, an eminent pharmacologist located in Philadelphia, PA at Thomas Jefferson University. Waldman is a relatively frequent name in the United States (with 208 researchers with an identical patronym in the AAMC faculty roster); the combination `"waldman s"` is common to 3 researchers in the same database. A simple search query for `"waldman sa"[au] OR "waldman s"[au]` returns 377 publications at the time of this writing. However, a more refined query, based on Professor Waldman's biosketch returns only 256 publications.[v]

The above example also makes clear how we deal with the issue of inconsistent publication names. PubHarvester gives the end-user the option to choose up to four *PubMed*-formatted names under which publications can be found for a given researcher. For example, Louis J. Tobian, Jr. publishes under `"tobian l"`, `"tobian l jr"`, and `"tobian lj"`, and all three names need to be provided as inputs to generate a complete publication listing. Furthermore, even though Tobian is a relatively rare name, the search query needs to be modified to account for these name variations, as in `("tobian l"[au] OR "tobian lj"[au])`.

[iv] `http://www.pubmed.gov/`

[v] `(((("waldman sa"[au] NOT (ether OR anesthesia)) OR ("waldman s"[au] AND (murad OR philadelphia[ad] OR west point[ad] OR wong p[au] OR lasseter kc[au] OR colorectal))) AND 1980:2013[dp])`

Appendix C: *PubMed* Related Citations Algorithm [PMRA]

Algorithm overview. The *PubMed* Related Citations Algorithm [PMRA] underlies the "related articles" search feature in *PubMed*. Lin and Wilbur (2007) develop a topic-based similarity model designed to help a typical user search through the literature by presenting a set of records topically related to a focal article returned by a *PubMed* search query.

Specifically, PMRA relies on Bayes' Theorem to estimates the probability that an individual is interested in document a given expressed interest in document b. They focus on the following relationship:

$$\Pr(a|b) \propto \sum_{j=1}^{N} \Pr(a|s_j)\Pr(b|s_j)\Pr(s_j),$$

where $\{s_1, ..., s_N\}$ denotes the entire set of mutually exclusive topics that could possibly be contained within a, b, or any other document of interest. Lin and Wilbur (2007) then make assumptions about the underlying arrival rates of terms within documents and how likely the occurrence of a term within a document actually reflects the true nature of that document. From these assumptions, the authors arrive at a topic weighting function, $w_{j,x}$, that describes how important a topic s_j is to any document x, and a document scoring function, $Sim(a, b)$, that quantifies the similarity between a and b, given by:

$$w_{j,x} = \lambda_{j,x} \times \sqrt{\frac{1}{f_j}}$$

$$Sim(a, b) = \sum_{j=1}^{N} w_{j,a} \times w_{j,b},$$

where f_j is the frequency of topic s_j in the entire corpus and $\lambda_{j,x}$ is based on a series of Poisson arrival rate parameters and the number of times topic s_j occurs within document x. Intuitively, two documents are more likely to be similar when they both use topics that are rare (f_j is low) many times ($\lambda_{j,x}$ is high). The authors estimate, optimize and experimentally confirm parameters to align with human assessments. They also report that one fifth of "non-trivial" browser sessions in *PubMed* invoke PMRA at least once, providing some "ground truth" for the view that the algorithm captures meaningful intellectual linkages between documents.

Defining topics. The algorithm relies on three types of text information to derive a list of potential topics: MeSH terms, abstract words, and title words. MeSH is the National Library of Medicine's [NLM] controlled vocabulary thesaurus. It consists of terms arranged in a hierarchical structure that permit searching at various levels of specificity (there are over 28,000 descriptors in the 2018 edition of MeSH). Almost every publication in *PubMed* is tagged with a set of MeSH terms (between 1 and 103 in the current edition of *PubMed*, with both the mean and median approximately equal to 11). NLM's professional indexers are trained to select indexing terms from MeSH according to a specific protocol, and consider each article in the context of the entire collection (Bachrach and Charen 1978; Névéol et al. 2010).

The presence of MeSH terms is crucial for the performance of the PMRA algorithm in two separate respects. Directly, because the MeSH terms are appended to the list of abstract words and title words to form the set of topics present in a *PubMed* record. Indirectly, because PMRA uses MeSH terms as informative markers to separate "elite" from "non-elite" topics in each record, and relies on a mixture of two Poisson distributions (one for elite terms, one for non-elite terms) to estimate the probability that a document is about a topic, given that we observe its corresponding term (abstract word, title word, MeSH term) a certain number of times in the document.

The reliance of PMRA on MeSH terms offers both advantages and disadvantages from the standpoint of our study. On the positive side of the ledger, professional indexers with domain expertise annotate articles with

MeSH terms—the authors are not involved. Professional annotators are probably less subject than authors to demand effects, whereby keywords are chosen endogenously to appeal to an audience of potential readers, referees, and journal editors. As such, they are relatively more stripped of "social baggage" than author-chosen keywords would be.[vi] Research in information science backs up the claim that MeSH terms can be seen as representing standardized and high-quality summaries of a particular publication (Bhattacharya et al. 2011).

On the negative side of the ledger, two features of the MeSH annotation process deserve mention. First, MeSH terms suffer from a keyword vintage problem as well as a left-censoring problem; these two problems are inextricably linked. Indexers may have available a lexicon of permitted keywords which is itself out of date. NLM continually revises and updates the MeSH vocabulary in an attempt to neutralize keyword vintage effects, but articles are not systematically backward-annotated. Take, for example the paper by Emmanuelle Charpentier and Jennifer Doudna which appeared in *Science* in June 2012 (Jinek et al., *"A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity"*) and established the viability of the CRISPR-Cas9 system for genome editing. The article is tagged by 11 unique MeSH terms, but CRISPR is not one them. This is of course because the CRISPR keyword was not part of the controlled MeSH thesaurus in 2012—it was "born" as a keyword in 2013!

Second, human indexers are not necessarily impervious to scientific fads and fashions. In their efforts to be helpful to *PubMed* users, they may use combinations of keywords that reflect the conventional views of the field. Probabilistic topic models such as PMRA assume that the scientific corpus has been correctly indexed. But what if the indexers who chose the keywords brought their own "conceptual baggage" to the indexing task, so that the pictures that emerge from this process are more akin to their conceptualization than to those of the scientists whose work it was intended to study? In our view, "indexer effects" (in the parlance of Whittaker 1989) present a more benign challenge. A number of studies have asked authors to validate *ex post* the quality of the keywords selected by independent indexers, with generally encouraging results (Law and Whittaker 1992). Inter-indexer reliability is also very high (Wilbur 1998).

There is an additional reason why these challenges deserve less emphasis than might appear at first blush, at least from the standpoint of accurately capturing intellectual relatedness. PMRA relies on abstract words and title words as well as MeSH terms. Going back to the Jinek et al. (2012) article, the word "CRISPR" appears four separate times in the abstract. PMRA can therefore link this foundational paper to 218 other articles, which will often be annotated with CRISPR-relevant MeSH terms (e.g., "CRISPR-Associated Proteins" or "CRISPR-Cas Systems.") In other words, the inclusion of title/abstract words help remedy unpleasant features of the MeSH annotation process. In so doing, however, they weaken our initial claim that the linkages revealed by PMRA are purely intellectual, devoid of "social baggage." For this reason, below we will explicitly look at the extent to which omitting abstract and title words from the input used by PMRA to generate the list of intellectual neighbors alters our benchmark set of results. Figure C1 depicts how the multiplier of the unconditional probability that two articles are related through PMRA is affected by the number of MeSH terms that overlap between the two records. For example, two articles picked at random are 255 times more likely to be related if they share 5 MeSH terms instead of only one. Note that the baseline unconditional probability that two articles are related when they share only one MeSH term is quite low, on the order of $1 \div 1,000,000$.

Implementation details. Using the MeSH keywords as input, PMRA essentially defines a distance concept in idea space such that the proximity between a source article and any other *PubMed*-indexed publication can be assessed. The following paragraphs were extracted from a brief description of PMRA:

> *The neighbors of a document are those documents in the database that are the most similar to it. The similarity between documents is measured by the words they have in common, with some adjustment for document lengths. To carry out such a program, one must first define what a word is. For us, a word is basically an*

[vi] Importantly, the assignment of MeSH keywords does NOT take into account references cited in the publication.

unbroken string of letters and numerals with at least one letter of the alphabet in it. Words end at hyphens, spaces, new lines, and punctuation. A list of 310 common, but uninformative, words (also known as stopwords) are eliminated from processing at this stage. Next, a limited amount of stemming of words is done, but no thesaurus is used in processing. Words from the abstract of a document are classified as text words. Words from titles are also classified as text words, but words from titles are added in a second time to give them a small advantage in the local weighting scheme. MeSH terms are placed in a third category, and a MeSH term with a subheading qualifier is entered twice, once without the qualifier and once with it. If a MeSH term is starred (indicating a major concept in a document), the star is ignored. These three categories of words (or phrases in the case of MeSH) comprise the representation of a document. No other fields, such as Author or Journal, enter into the calculations.

Having obtained the set of terms that represent each document, the next step is to recognize that not all words are of equal value. Each time a word is used, it is assigned a numerical weight. This numerical weight is based on information that the computer can obtain by automatic processing. Automatic processing is important because the number of different terms that have to be assigned weights is close to two million for this system. The weight or value of a term is dependent on three types of information: 1) the number of different documents in the database that contain the term; 2) the number of times the term occurs in a particular document; and 3) the number of term occurrences in the document. The first of these pieces of information is used to produce a number called the global weight of the term. The global weight is used in weighting the term throughout the database. The second and third pieces of information pertain only to a particular document and are used to produce a number called the local weight of the term in that specific document. When a word occurs in two documents, its weight is computed as the product of the global weight times the two local weights (one pertaining to each of the documents).

The global weight of a term is greater for the less frequent terms. This is reasonable because the presence of a term that occurred in most of the documents would really tell one very little about a document. On the other hand, a term that occurred in only 100 documents of one million would be very helpful in limiting the set of documents of interest. A word that occurred in only 10 documents is likely to be even more informative and will receive an even higher weight.

The local weight of a term is the measure of its importance in a particular document. Generally, the more frequent a term is within a document, the more important it is in representing the content of that document. However, this relationship is saturating, i.e., as the frequency continues to go up, the importance of the word increases less rapidly and finally comes to a finite limit. In addition, we do not want a longer document to be considered more important just because it is longer; therefore, a length correction is applied.

The similarity between two documents is computed by adding up the weights of all of the terms the two documents have in common. Once the similarity score of a document in relation to each of the other documents in the database has been computed, that document's neighbors are identified as the most similar (highest scoring) documents found. These closely related documents are pre-computed for each document in PubMed so that when one selects Related Articles, the system has only to retrieve this list. This enables a fast response time for such queries.[vii]

For a given source article, PMRA yields the following output: (i) an ordered list of intellectually related articles with a fixed length; (ii) a cardinal measure of distance between the source and each related article, which we have normalized such that a source is always 100% related to itself, and relatedness decreases as one goes down the ranking of the ordered list of neighbors.

Cutoff Rules. The algorithm uses a cutoff rule to determine the number of related articles associated with a given source article. First, the 100 most related records by similarity score are returned. Second, a reciprocity rule is applied to this list of 100 records: if publication x is related to publication y, publication y must also be related to publication x. As a result, there is no fixed number of related articles for a source article. On the contrary, the total number of related articles can be of arbitrary large size, and certainly much higher than 100. Figure C2, Panel A displays the histogram for the distribution of the number of related articles for the 35,409 source articles in our main sample. The mean number of articles is 153 and the median 119. Surprisingly, however, 25% or so of the source articles have *less* than 100 related articles associated with them. In part, this is an artefact of some data construction choices, as we eliminate related articles outside the [1965; 2006] date range, or related articles that are not original articles (reviews, editorials, etc.), or related articles in journals not indexed by the *Web of Science*. And yet, even after accounting for these

[vii] Available at `http://ii.nlm.nih.gov/MTI/related.shtml`

factors, slightly more than 10% of the source articles have less than 100 intellectual neighbors, which is surprising given the documented cutoff rule whereby PMRA supposedly always starts from a list of 100 neighbors, and then possibly add to this list via symmetry.

We investigated this peculiar feature of the data; PMRA appears to have a second cutoff rule based on the cardinal relatedness score. For each source article, we computed the minimum relatedness score, and graphed the resulting distribution (Figure C2, Panel B). One can observe a mass point around 0.10 (corresponding to 3% of the source articles), meaning that PMRA will fail to expand the set of neighbors all the way up to 100 articles if it finds out that doing so would mean including related articles with relatedness < 0.10.[viii]

The presence of this second cutoff is in an important respect a welcome (if idiosyncratic and poorly documented) feature of the algorithm. If the cutoff was downward-rigid at 100, then after a star scientist had passed away, PMRA would need to reach into a set of articles that are in fact quite intellectually distant from the source to fill the void mechanically induced by the fact that the deceased star cannot contribute to his own subfields. Figure C2, Panel C confirms that it is not the case. It depicts, for both treated and control source articles, the distribution of relatedness score for the least related article associated with each source article, only taking into account the articles written *after* the death (or counterfactual death) of the star. The two distributions are quite close to one another; if anything, there are slightly more control source articles that lie at the cardinal cutoff value of 0.10, relative to treated source articles. In other words, we find no evidence of "overexpansion" in less proximate intellectual domains for treated fields, relative to control fields, in the period that follows the death of an eminent scientist.

One final check is to look for stability over time, both for the ordinal cutoff and the cardinal cutoff. A maintained assumption for our research design is that these cutoffs do not vary over time differentially for treated and control fields. We investigate cutoff stability by running a regression of each subfield's log size (respectively, each subfield's log odds of the lowest relatedness score) onto journal effects, number of authors effects, 36 source publication year effects (from 1967 to 2002, 1966 is the omitted variable), and 36 source publication year by treatment status interaction terms. We graph the coefficient estimates corresponding to these interaction effects on Figure C3, which are for the most part imprecisely estimated zeros, and do not exhibit any specific upward or downward trend. From all these analysis, we conclude that there is no reason to suspect that PMRA's cutoff rules impact treated and control source articles in a differential way.

From source article to subfield: An Example. Given our set of source articles, we delineate the scientific fields to which they belong by focusing on the set of articles returned by PMRA that satisfy three additional constraints: (i) they are original articles (as opposed to editorials, comments, reviews, etc.); (ii) they were published in or before 2006 (the end of our observation period); and (iii) they appear in journals indexed by the *Web of Science* (so that follow-on citation information can be collected). In Figure C4, we illustrate the use of PMRA with an example taken from our sample. Consider *"The transcriptional program of sporulation in budding yeast"* (*PubMed* ID #9784122), an article published in the journal *Science* in 1998 originating from the laboratory of Ira Herskowitz, an eminent UCSF biologist who died in 2003 from pancreatic cancer. PMRA returns 72 original related journal articles for this source publication.[ix] Some of these intellectual neighbors appeared before the source to which they are related, whereas others were published after the source. Some represent the work of collaborators, past or present, of Herskowitz's, whereas others represent the work of scientists in his field he may never have come in contact with during his life, much less collaborated with. The salient point is that nothing in the process through which these related articles are identified biases us towards (or away from) articles by collaborators, frequent citers of

[viii] There is a smattering of source articles for which the minimum relatedness is below 0.10. Upon closer examination, these source articles have no abstracts in *PubMed*, or do not have MeSH terms available. We investigated the sensitivity of our main results to dropping these subfields from the analysis (Appendix E, Table Ex).

[ix] Why exactly 72? In fact, PMRA lists 152 "intellectual neighbors" for PubMed ID 9784122. But once we exclude articles published after 2006 (the end of our observation period), purge from the list reviews, editorials and other miscellaneous "non-original" content, and drop a handful of articles that appeared in minor journals not indexed in Thomson-Reuter's *Web of Science*, the number of publications associated with this source article indeed drops to 72.

Herskowitz's work, or co-located researchers. Rather, the only determinants of relatedness are to be found in the overlap in MeSH keywords between the source and its potential neighbors.

PubMed ID #9784122 appeared in the October 23^{rd} 1998 issue of the journal *Science* and lists 15 MeSH terms and 5 substances. Consider now its second most-related (listed in Figure C1), *PubMed* ID #12242283 *"Phosphorylation and maximal activity of Saccharomyces cerevisiae meiosis-specific transcription factor Ndt80 is dependent on Ime2."* It appeared in *Molecular and Cell Biology* in October of 2002 and has 24 MeSH terms (resp. 11 substances). Figure C5 displays the MeSH terms that tag this article along with its source *PubMed* ID #9784122. The keywords that overlap exactly have been highlighted in dark blue; those whose close ancestors in the MeSH keyword hierarchical tree overlap have been highlighted in light blue. These terms include common terms such as `Saccharomyces cerevisiae` and `Transcription Factors` as well as more specific keywords including `NDT80 protein, S cerevisiae` and `Gene Expression Regulation, Fungal.`

PMRA also provides a cardinal dyadic measure of intellectual proximity between each related article and its associated source article. In this particular instance, the relatedness score of "Phosphorylation..." is 94%, whereas the relatedness score for the most distant related article in Figure C4, "Catalytic roles of yeast..." is only 62%.

Delineating subfields. In the five years prior to his death (1998-2002), Herskowitz was the last author on 12 publications, the publications most closely associate with his position as head of a laboratory. For each of these source publications, we treat the set of publications returned by PMRA as constituting a distinct subfield, and we create a subfield panel dataset by counting the number of related articles in each of these subfields in each year between 1975 and 2006.

An important characteristic of the subfields subfields generated by this procedure is that they correspond to quite compact intellectual neighborhoods. One window into the extent of intellectual breadth for PMRA-generated subfields is to gauge the overlap between the articles that constitute any pair of subfields associated with the same star. In the sample, the 452 deceased stars account for 3,076 subfields, and 21,661 pairwise combination of subfields (we are only considering pairs of subfields associated with the same individual star). Figure C6 displays the histogram for the distribution of overlap, which is extremely skewed. A full half of these pairs exhibit exactly zero overlap, whereas the mean of the distribution is 0.06. To find pairs of subfields that display substantial amounts of overlap (for example, half of the articles in subfield 1 also belong in subfield 2), one must reach far into the right tail of the distribution, specifically, above the 98^{th} percentile.

Given a source article published in year t, PMRA will tend to find the largest number of neighbors contemporaneously, slightly fewer neighbors–but still a large proportion of them all–during years $t-1$ and $t+1$, a slightly lower number still in years $t-2$ and $t+2$, etc. In other words, PMRA creates lists of intellectual neighbors such that, when rolled up at the year level, will generate subfields whose life cycle has an inverted U-shape, with the peak of the U corresponding to the year of publication for the source. This does not strike us as an implausible feature of the scientific process: papers related to a focal one will be more likely to appear in close temporal proximity with it. Importantly, this feature of PMRA affects treated and control subfields in a precisely symmetric fashion.

To illustrate this empirically, we took a random sample of 5,000 articles in *PubMed* (original articles, in journals indexed by web of science, that appeared between 1965 and 2003–the same range of years as for our source articles) and computed the average number of articles entering those subfields in a range of $[-10; +10]$ years after the publication of the source. This yields a pronounced inverted-U shape, as seen on Figure C7. Interestingly, the decay in the outer years is not symmetric: PMRA finds more neighbors in the future than in the past. This may reflect the steadily expanding universe of publications, such that there will mechanically be more candidates to be included as related neighbors going forward in time, relative to going backward in time. The same tendency would of course apply equally to control and treated subfields.

Robustness checks. The production version of PMRA is used by thousands of scientists every day to assist their search of the biomedical literature. The foregoing discussion has shown that some idiosyncrasies baked into the algorithm are not necessarily desirable from a research standpoint. How would our benchmark set of results change, for instance, if the subfields were expanded in size? Or if a cardinal cutoff rule determined the boundary of a subfield? Or if only MeSH terms, rather than the combination of MeSH terms and abstract/title words, were used to assess the similarity between the documents in a subfield? Below, we avail ourselves to an off-line version of PMRA that was explicitly built to allow some limited experimentation with featured of the PMRA algorithm.[x] Using this software tool, we can generate the relatedness score between a source article in *PubMed* and a string of text. We manipulate that string of text to generate relatedness scores between our source articles and an expanded set of candidate related articles under different scenarios.

Before doing so, however, we need to create an expanded list of "candidate" related articles, because we lack the computing power to check each source article against the entire *PubMed* corpus.[xi] Our approach is to combine the related articles (denoted $PMRA^1$ articles below) with the related articles of the related articles (denoted $PMRA^2$ articles below) as the candidate set. Using the cardinal relatedness score generated by the off-line, tunable version of the software, we then use a simple cutoff rule to delineate the expanded subfields: we retain only those articles with cardinal relatedness score greater than 0.20 (the median). In addition, as is the case for the benchmark set of subfields, we also eliminate non-original articles, articles that fall outside of our date range, articles not written in English, and articles that appear in journals not indexed by the *Web of Science*. We repeat this exercise, except that we set loose the tunable version of PMRA on candidate related articles that are summarized solely by their MeSH terms (i.e., abstract/title words are not taken into account).

Figure C8 displays the histogram of the distribution for subfields constructed using this novel set of rules. The mean stands at 891 articles, the median at 625 articles, with a maximum value of 7,112. These subfields are therefore much larger than those generated by the production version of PMRA. Table C1 replicates our benchmark set of specifications (columns 1, 2, and 3 of Table 3) on these new data. The leftmost three columns correspond to the version where abstract/title words and MeSH terms are used to calculate relatedness score; the rightmost three columns correspond to the version where the input into the calculation of relatedness is limited to the MeSH terms. The magnitudes of the effects are a bit larger than those observed in Table 3; the coefficients are also more precisely estimated. Figure C9 replicates Panel C of Figure 2 on the new data. Panel A of Figure C9 corresponds to a dynamic version of the specification in column (3) of Table C1, whereas Panel B of Figure C9 corresponds to a dynamic version of the specification in column (6) of Table C1. In both of these pictures, there appear to be a slight pre-trend in that activity in the field picks up slightly before the death of the star scientist. The magnitudes, however, are very small, marginally significant, and substantially smaller than those found in the post-death period, providing reassurance regarding the robustness of our core results.

[x] We thank Kyle Myers from the NBER for graciously allowing us access to this software, which forms the basis of his manuscript entitled "The Elasticity of Science" (Myers 2018). Note that it relies on a version of *PubMed* that is not complete—about 10% of the online version of the database have no counterparts in the off-line version, but these articles appear to be missing at random.

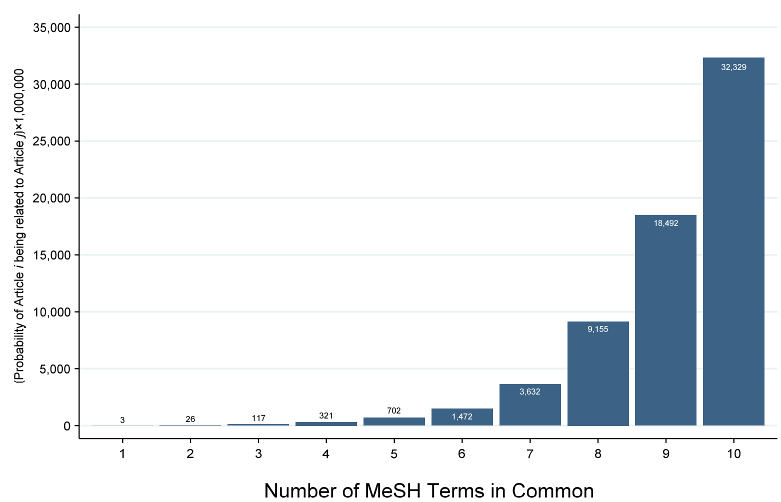
[xi] There would be close to half a trillion article pairs to check, even after eliminating articles outside of our date range, non-original content, articles in other languages, etc.

Table C1: Alternate Subfield Definitions

	Expanded Neighborhoods			Expanded Neighborhoods, MeSH Terms Only		
	All Authors	Collabs. Only	Non-Collabs. Only	All Authors	Collabs. Only	Non-Collabs. Only
After Death	0.098**	-0.321**	0.120**	0.071**	-0.327**	0.089**
	(0.026)	(0.047)	(0.026)	(0.022)	(0.045)	(0.022)
Nb. of Investigators	6,237	6,194	6,237	6,226	6,189	6,226
Nb. of Fields	33,987	33,732	33,981	33,928	33,761	33,928
Nb. of Field-Year Obs.	1,390,415	1,380,078	1,390,169	1,398,549	1,391,664	1,398,549
Log Likelihood	-5,918,924	-1,508,675	-5,704,068	-8,106,163	-1,818,687	-7,895,247

Note: Estimates stem from conditional (subfield) fixed effects Poisson specifications. The dependent variable is the total number of publications in a subfield in a particular year (similar to Table 3, columns 1 through 3). All models incorporate a full suite of year effects and subfield age effects, as well as a term common to both treated and control subfields that switches from zero to one after the death of the star, to address the concern that age, year and individual fixed effects may not fully account for trends in subfield entry around the time of death for the deceased star. The first three columns use subfields that comprise both $PMRA^1$ and $PMRA^2$ articles, but where the input data includes abstract/title words plus MeSH terms, just as in the production version of the algorithm. In contrast, in the second set of three columns, subfields have been constructed while ignoring abstract/title words for the candidate related articles. Robust standard errors in parentheses, clustered at the level of the star scientist. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$.

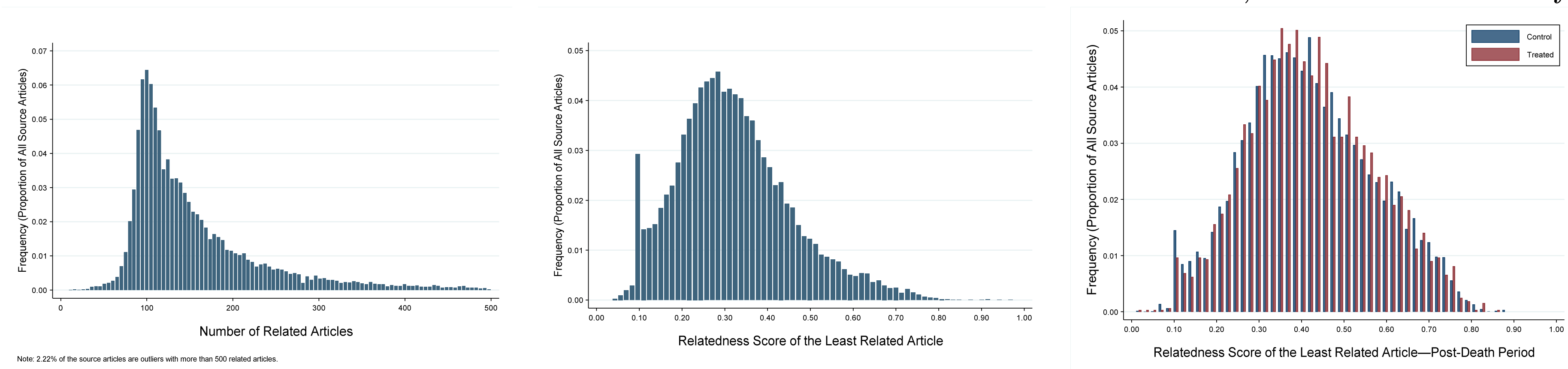
Figure C1: MeSH Term Overlap & Relatedness



Note: This figure depicts the relationship between MESH term overlap and being classified as related by PMRA based on a random sample of approximately 130 million article pairs in PubMed (formed from a random sample of 15,400 individual articles). With exactly one MeSH term in common, the base probability of being related is on the order of 1/1,000,000. That probability increases extremely steeply as the number of MeSH terms shared between any two random articles moves beyond 4 terms in common.

Figure C2
Subfield Size and PMRA Cutoff Rules

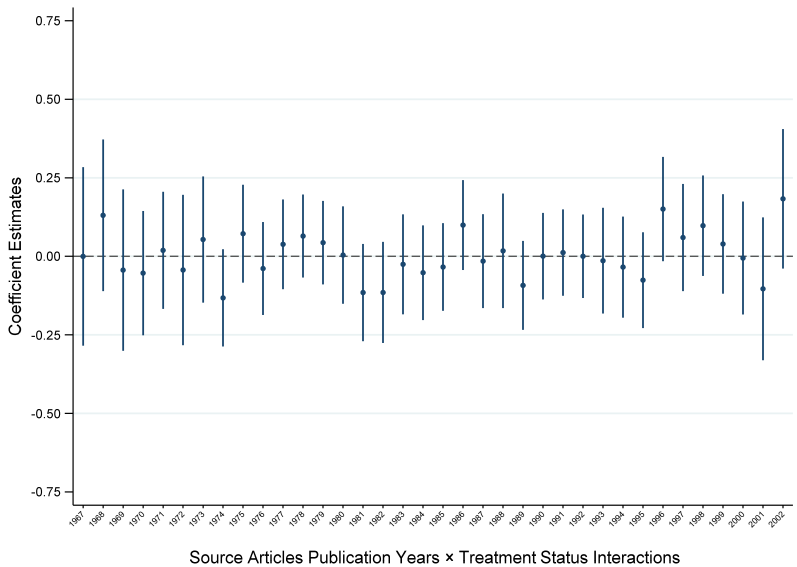
A. **Ordinal Cutoff** | **B. Cardinal Cutoff** | **C. Cardinal Cutoff, by Treatment Status, Post-Death Period Only**



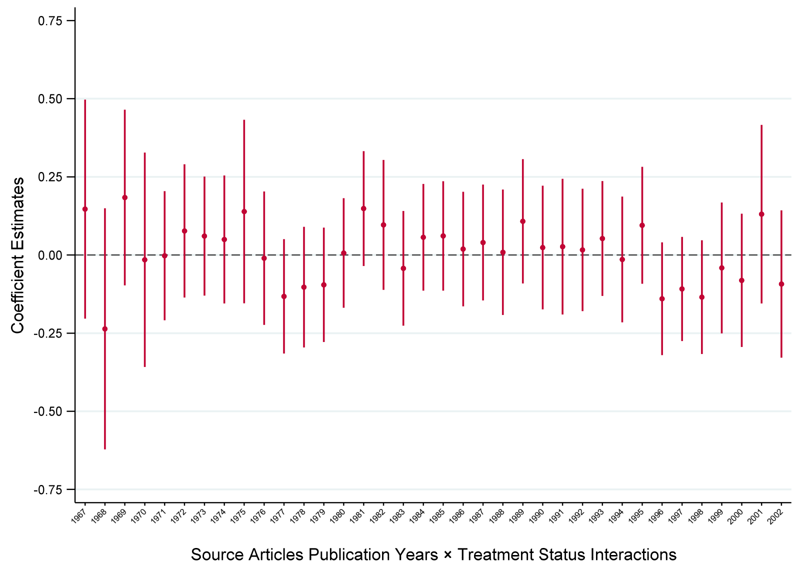
<u>Note</u>: We document the rules that govern the cutoff in the number of related articles associated with each source. Panel A depicts the histogram for the distribution of related articles after filtering out "undesirable" publications (such as reviews and other non-original material, non-English publications, etc.). Panel B depicts the distribution of the relatedness score for the least related article associated with each source article in our data. There is a mass point at 0.10 that corresponds to an additional cutoff rule in PMRA. A smattering of source publications have some related articles with relatedness score below 0.10, but the overwhelming majority of those are incomplete records: missing abstract, missing MeSH terms, or both. These account for less than 0.5% of the source articles. Finally, Panel C compares the relatedness of the least related article for each source, by treatment status, and solely for the related articles that appeared after the death (respectively counterfactual death) of a star.

Figure C3
Temporal Stability of Cutoff Rules

A. Ordinal Cutoff (Subfield Size)



B. Cardinal Cutoff (Lowest Relatedness Score)



Note: We regress the log of pre-death subfield size (Panel A) and the log odds of the relatedness score for the least related article (Panel B) onto (i) journal fixed effects; (ii) a suite of indicator variables for the source article's number of authors; (iii) source article year of publication effects; and (iv) interaction terms between each year of publication and a treatment status indicator. The graphs report the coefficient estimates, along with their associated 95% confidence interval (corresponding to robust standard errors, clustered at the level of the star) for these 36 interaction terms.

Figure C4: From Source to Related Articles

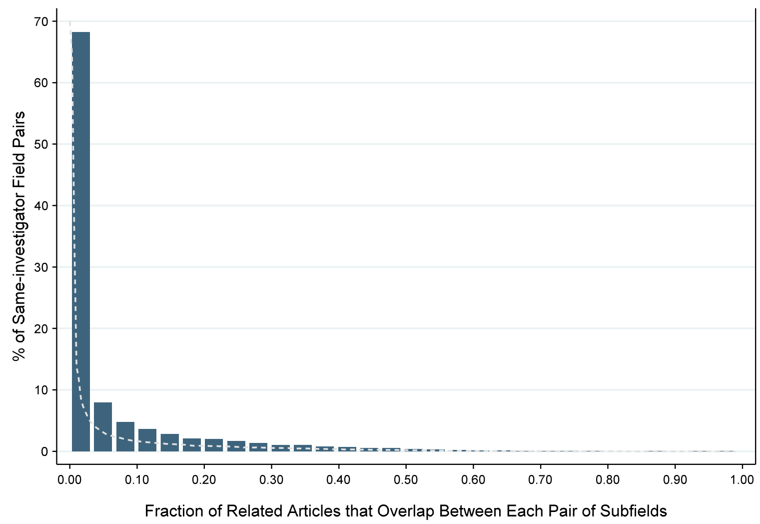
Note: We illustrate the process of identifying the related articles through the use of an example. Ira Herskowitz, a superstar scientist in our sample, died in 2003. In the five years prior to his death (1998-2002), Herskowitz was the last author on 12 publications. One of these publications is *"The transcriptional program of sporulation in budding yeast,"* an article published in the journal *Science* in 1998. On the right-hand side panel, one sees that PMRA identifies 72 related articles related to this source publication. Each of these related articles can then be parsed in a variety of ways. In particular, their authorship list can be matched to the AAMC Faculty Roster, which allows us to distinguish between collaborators of Herskowitz's and non-collaborators, as well as between the subfield's insiders vs. outsiders. Eight out of the 72 articles have a former or current collaborator on the authorship roster. Twenty two of the 72 articles in the subfield cite the source article, while the source articles references eight of the articles in the subfield.

Figure C5: PMRA and MeSH Term Overlap—An Example

Source Article	**PMRA-Linked Article**
Chu et al., "The transcriptional program of sporulation in budding yeast." *Science*, 1998.	Sopko et al., "Phosphorylation and maximal activity of Saccharomyces cerevisiae meiosis-specific transcription factor Ndt80 is dependent on Ime2." *MCB*, 2002.
PMID #9784122	**PMID #12242283**
MeSH Terms	**MeSH Terms**
Animals	Active Transport, Cell Nucleus
Chromosomes, Fungal	Binding Sites
DNA-Binding Proteins*	Cell Cycle Proteins*
Fungal Proteins	Cell Nucleus
Gene Expression Regulation, Fungal*	DNA-Binding Proteins*
Genes, Fungal	Fungal Proteins*
Genome, Fungal	Gene Expression Regulation, Fungal*
Humans	Genes, Fungal
Meiosis	Intracellular Signaling Peptides and Proteins
Morphogenesis	Meiosis*
Organelles	Phosphorylation
Saccharomyces cerevisiae*	Promoter Regions, Genetic
Spores, Fungal	Protein Kinases*
Transcription Factors	Protein-Serine-Threonine Kinases
Transcription, Genetic*	Recombinant Fusion Proteins
	Saccharomyces cerevisiae
	Saccharomyces cerevisiae Proteins*
	Spores, Fungal
	Substrate Specificity
	Transcription Factors*
	Transcriptional Activation
Substances	**Substances**
DNA-Binding Proteins	Cell Cycle Proteins
Fungal Proteins	DNA-Binding Proteins
NDT80 protein, S cerevisiae	Fungal Proteins
Saccharomyces cerevisiae Proteins	Intracellular Signaling Peptides and Proteins
Transcription Factors	NDT80 protein, S cerevisiae
	Recombinant Fusion Proteins
	Saccharomyces cerevisiae Proteins
	Transcription Factors
	Protein Kinases
	IME2 protein, S cerevisiae
	Protein-Serine-Threonine Kinases

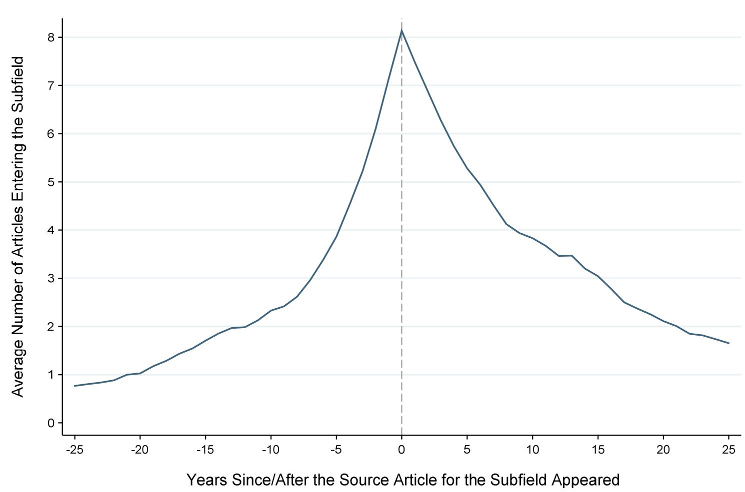
Note: We compare the MeSH terms for the number of MeSH terms for the source article in Figure C4, along with those of its most proximate intellectual neighbor according to PMRA.

Figure C6
Article Overlap Between Subfield Pairs



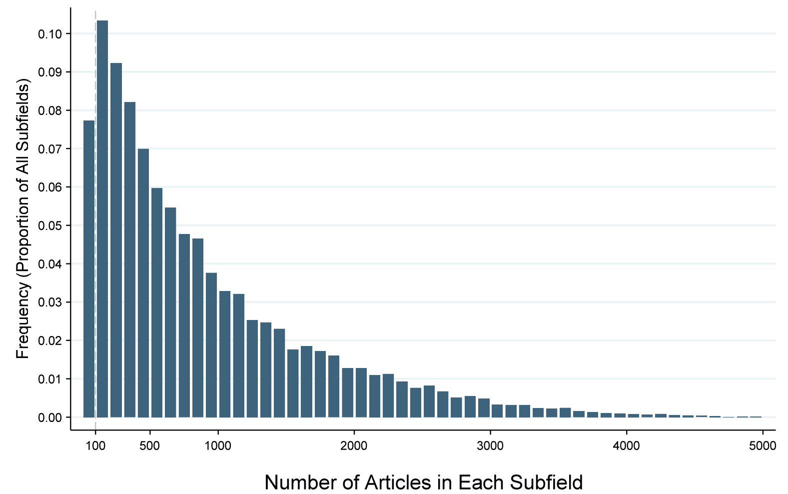
Note: We compute the share of related articles that are shared between pairs of PMRA-delineated subfields. To be conservative, we focus the analysis on 21,661 subfield pairs where a deceased superstar was the last author on both of the associated source articles.

Figure C7
Distribution of Activity in Subfields Over Time



Note: This figure illustrates the timing of articles entering the subfields for a random sample of 5,000 articles in PubMed (original articles, in journals indexed by *Web of Science*, that appeared between 1965 and 2003—the same range of years as for the source articles in our analytic sample), and we run them through PMRA, rolling up the count of articles up to the subfield-year level (as in our regressions).

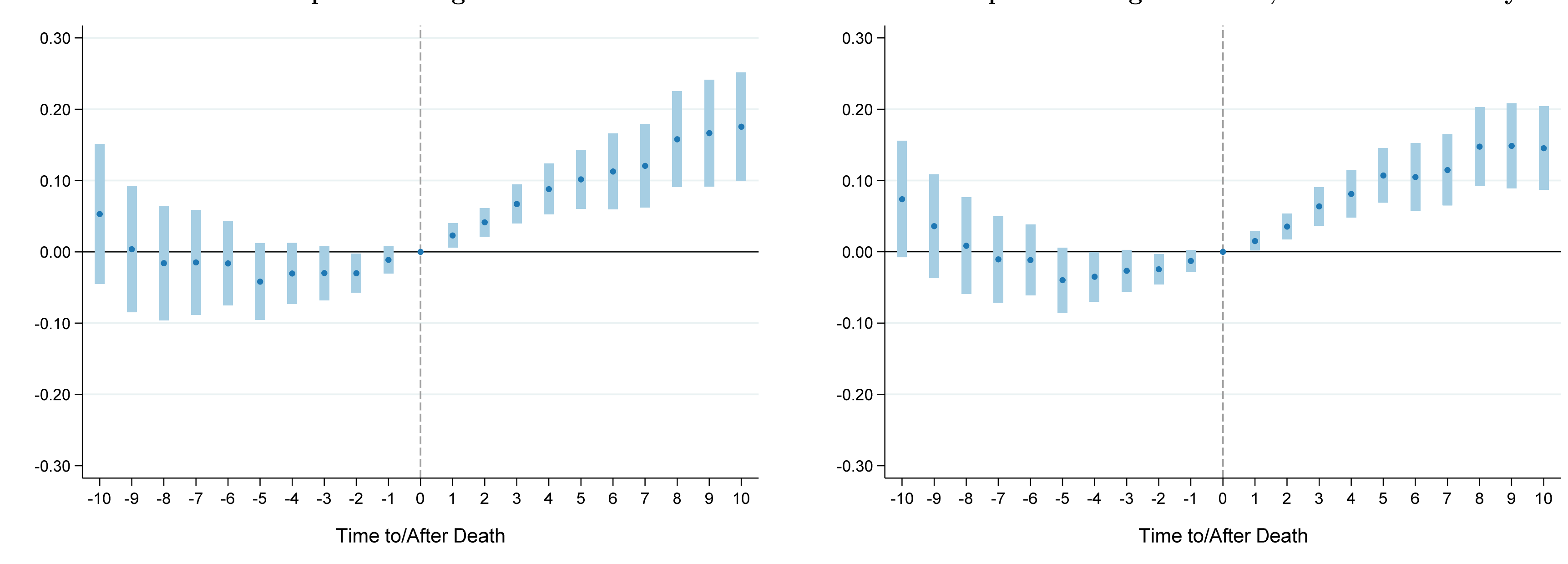
Figure C8
Distribution of Expanded Neighborhood Subfield Size



Note: The articles that are candidate for membership in each subfield satisfy the following conditions: $PMRA^1$ or $PMRA^2$. We then compute relatedness in this expanded neighborhood using the tunable version of PMRA. We discard every article with new relatedness score less than 0.20 (the median in the sample). As a result, there is a cardinal cutoff, but no ordinal cutoff that delineates subfield boundaries. 35 (0.1%) of the fields are outliers with more than 5,000 articles. In the histogram above, we make use of abstract & title words, in addition to MeSH terms, to assess relatedness through PMRA.

Figure C9
Dynamics of Subfield Entry—Non Collaborators
Alternate Subfield Definitions

A. Expanded Neighborhoods **B. Expanded Neighborhoods, MeSH Terms Only**



Note: The dark blue dots in the above plots correspond to coefficient estimates stemming from conditional (subfield) fixed effects Poisson specifications in which publication flows in subfields are regressed onto year effects, subfield age effects, as well as 20 interaction terms between treatment status and the number of years before/after the death event (the indicator variable for treatment status interacted with the year of death is omitted). The specifications also include a full set of lead and lag terms common to both the treated and control subfields to fully account for transitory trends in subfield activity around the time of the death. The 95% confidence interval (corresponding to robust standard errors, clustered around star scientist) around these estimates is plotted with vertical light blue lines; Panel A corresponds to a dynamic version of the specification in the third column of Table C1; Panel B corresponds to a dynamic version of the specification in the sixth column of Table C1.

Appendix D: Construction of the Control Group

We detail the procedure implemented to identify the control subfields that help pin down the life-cycle and secular time effects in our difference-in-differences (DD) specification. Happenstance might yield a sample of stars clustered in decaying scientific fields. More plausibly, activity in the typical subfield might be subject to idiosyncratic life-cycle patterns, with their productive potential first increasing over time, eventually peaking, and thereafter slowly declining. Relying solely on subfields treated earlier or later as an implicit control group raises the worry that these time-varying omitted variables will not be fully captured by subfield age controls, particularly since dating the birth of a subfield is a process fraught with hazards.

To address this concern, we create an additional level of difference by selecting control subfields. Recall that selecting a subfield in our framework is akin to first selecting a source article and then using PMRA to harvest all the related articles to this source in intellectual space. Since the second step is fully automated, only the first step is really of concern. Practically, we will recruit control source articles from the set of articles authored by star scientists who do not die prematurely. But what makes a satisfactory control group? It is important to distinguish between *ex ante* vs. *ex post* criteria. *Ex ante*, one would like control source articles to have the following properties:

1. to be published contemporaneously with the source article for the treated subfield;
2. to be unrelated (in both an intellectual and a social sense) to the source article for the treated subfield;
3. to be of similar expected impact and fruitfulness, relative to the source article for the treated subfield;
4. to have a similar number of authors as the source article for the treated subfield;
5. to have a superstar author in the same authorship position and of approximately the same age as that occupied by the deceased superstar on the authorship roster of the source article for the treated subfield.

Ex post, it will be important for the control subfields to satisfy an additional condition: the treated and control subfields should exhibit very similar trends in publication activity and funding flows up to the year of treatment (i.e., the year of death for the treated superstar).

Coarsened Exact Matching. To meet these goals, we implement a “Coarsened Exact Matching” (CEM) procedure (Blackwell et al. 2009). The first step is to select a relatively small set of covariates on which we need to guarantee balance *ex ante*. This choice entails judgement, but is strongly guided by the set of criteria listed above. The second step is to create a large number of strata to cover the entire support of the joint distribution of the covariates selected in the previous step. In a third step, each observation is allocated to a unique strata, and for each observation in the treated group, control observations are selected from the same strata.

The procedure is coarse because we do not attempt to precisely match on covariate values; rather, we coarsen the support of the joint distribution of the covariates into a finite number of strata, and we match a treated observation if and only if a control observation can be recruited from this strata. An important advantage of CEM is that the analyst can guarantee the degree of covariate balance *ex ante*, but this comes at a cost: the more fine-grained the partition of the support for the joint distribution (i.e., the higher the number of strata), the larger the number of unmatched treated observations.

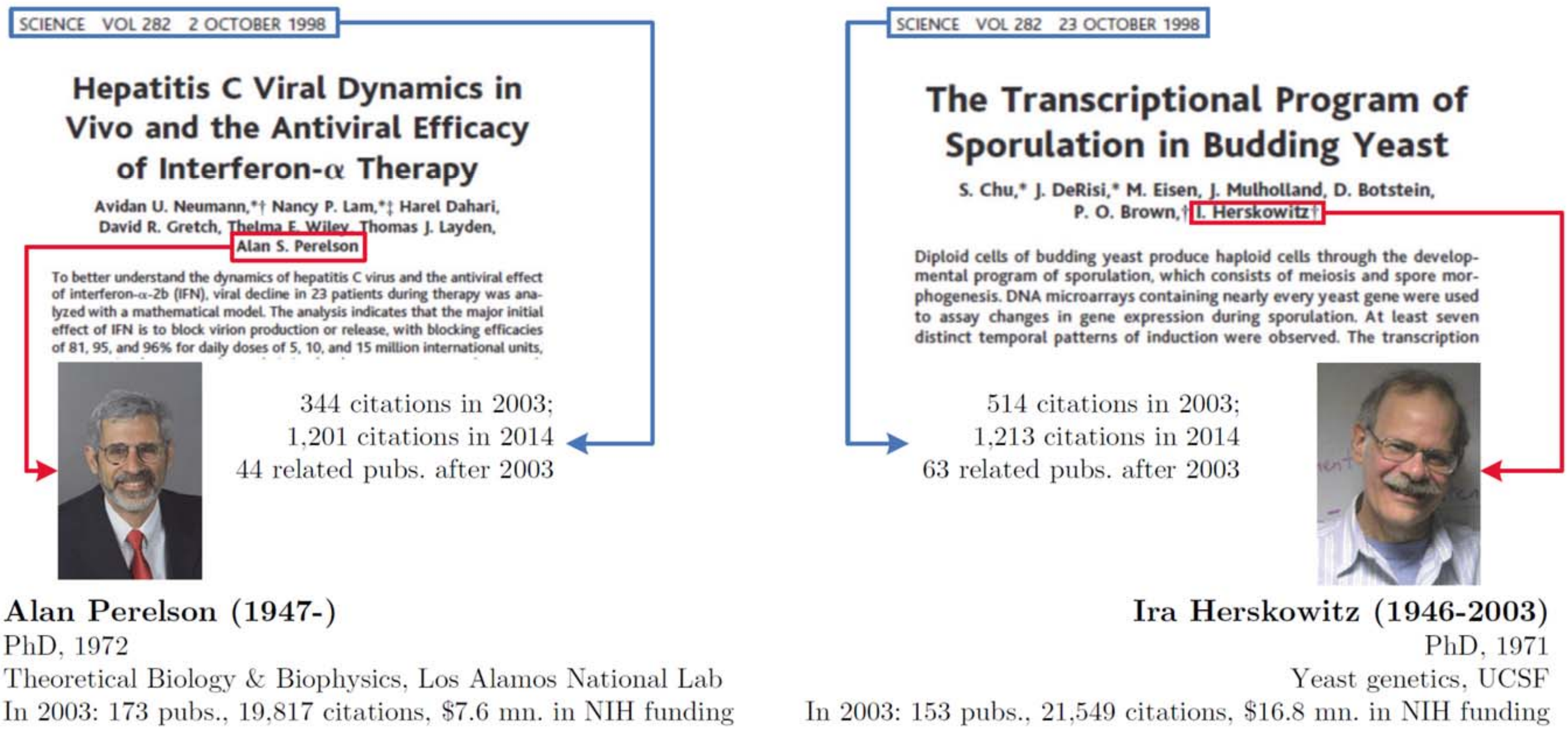
Implementation. We identify controls based on the following set of covariates (t denotes the year of death): star scientist career age; citations received by the article up to year t; number of authors; position of the star author on the authorship roster (only last authorship position is considered); journal; and year of publication. The first three covariates only need to match within relatively coarse bins. For instance, we create nine career age categories: less than 10 years; between 10 and 20 years; between 20 and 25 years; between 25 and 30 years; between 30 and 35 years; between 35 and 40 years; between 40 and 45 years; between 45 and 50 years, over 50 years of career age. Similarly, we coarsen the distribution of citations at baseline into five mutually exclusive bins: zero citations; between one and 10 citations; between 10 and 50 citations; between 50 and 120 citations; and more than 120 citations. In contrast, we impose an exact match on journal, publication year, and the star's authorship position.

We match approximately 75% of the treated source articles in this way. Some further trimming of the control articles is needed. First, we eliminate any control that shares any author with the treated source. Second, we eliminate any control article with a dead star scientist on its authorship roster, even if he appears in an intermediate position in the authorship list. Third, we drop every control that also happens to be related intellectually to its source as per PMRA. Finally, we drop from the data any source article that finds itself an orphan (i.e., not paired with any control) at the conclusion of this process. Figure D1 provides an illustrative example.

The final sample has 3,074 treated source articles and 31,142 control source articles. As can be seen in Figure D2, the distribution of activity levels, measured by cumulative publications up to the baseline year, is very similar between treated and control subfields. As well, there is no evidence of preexisting trends in activity, as demonstrated by the coefficient estimates graphed in Figure 1 and E1. In Table 2, treated and control subfields are very well-balanced on the covariates that formed the basis of the CEM matching procedure. This is true almost by construction. What is more surprising (and also welcome) is that the procedure balances a number of covariates that were not used as inputs for matching, such as various metrics of star eminence. For other covariates, we can detect statistically significant mean differences, though they do not appear to be substantively meaningful (e.g., 6.7% of control stars vs. 9.9% of treated stars are female).

Sensitivity Analyses. Human judgement matters for the outcome of the CEM procedure insofar as one must draw a list of "reasonable" covariates to match on, as well as decide on the degree of coarsening to impose. We have verified that slight variations in the implementation (e.g., varying slightly the number of cutoff points for the stock of baseline citations for the source; focusing on birth age as opposed to career age for the stars) have little impact on the main results.

Figure D1: Matching Procedure to Identify Controls for the Source Articles



Note: The two articles above illustrate the Coarsened Exact Matching (CEM) procedure. These two articles appeared in the journal *Science* in 1998. They received a similar number of citations up to the end of the baseline year (2002, one year before Herskowitz's death): 514 citations for Chu et al., 344 citations for Neumann et al. Note that Alan Perelson and Ira Herskowitz are both in last authorship position. They also obtained their PhD within a year of each other.

Appendix E: Extensions

Extended descriptive statistics. For space reasons, Table 2 provided descriptive statistics at baseline for only a selected set of right-hand side covariates and outcome variables. In Tables E1 and E2, we present descriptive statistics and correlation matrices for all the covariates and outcome variables that appear either in the main body of the manuscript, or in Appendixes E and F. Table E1a highlights balance between control and treated subfields at baseline for a simple transformation of the outcome variables. Recall that our outcome variables are of the form "number of articles in subfield i and year t that satisfy some condition," where examples of such conditions include, *inter alia, "by non-collaborators only, where these related authors had no prior participation in the subfield"* or *"by non-collaborators only, where the focal star is not cited in the list of references."* We transform these flow variables into cumulative stock variables, taking into account the years between the birth of the subfield and the year of death (or counterfactual death). So, for example, at baseline, the stock or related articles by non-collaborators that list references only outside the subfield are balanced between control and treated subfields (13.764 vs. 13.789).[xii]

Table E1b provides descriptive statistics for star-level (e.g., cumulative NIH funding at baseline) and subfield-level (e.g., commitment of the star to the subfield) covariates. These covariates are used to realize sample splits around their medians in Tables 6 and 7 of the manuscript, and in Tables E5, E7, and E8 of Appendix E. In Table E2, we also display correlation matrices for these variables. To make the matrix legible, we place correlations for subfield-level covariates and star-level covariates in separate tables (E2a and E2b). The correlations are typically reassuringly high across measures within a construct (e.g., , but low across constructs.

Event study graphs using the raw data. Figure E1 provides graphical evidence of the effect of star death on subfield entry using raw data. This involves an important simplification—anchoring the comparison between control and treated subfields on "experimental time" (the number of years elapsed since treatment), ignoring the fact that our death events are staggered over a long time period (1975 to 2003). Yet, these graphs provide visual evidence that the main effects of death on subfield growth or decline we document in regression specifications saturated with calendar year and age effects (Figure 2) are also apparent in the raw data. The graphs in Figure E1 also make vivid the life-cycle of subfields. Given a particular source article, PMRA creates a list of intellectual neighbors that, when added together at the year level, generate subfields whose evolution over time follows an inverted U-shape, with the peak of the U corresponding to the year of publication for the source.[xiii] Of course, these life cycle patterns are a reflection of design choices for PMRA. That being said, a plausible feature of the scientific process is that papers related to a focal one will be more likely to appear in close temporal proximity with it.

Sudden vs. Anticipated Death Events. To gain statistical power, our main results pool the subfields of stars who died suddenly with those of stars whose untimely passing was anticipated. Yet, the case for the exogeneity of a death event is stronger when it is sudden; when the death can be anticipated, it is theoretically possible for the star to engage in "intellectual estate planning," whereby particular scientists (presumably close collaborators) are anointed as exemplars of the next generation of leaders in the subfield. Table E3 breaks down our core set of results by cause of death, focusing on entry by non-collaborators only. Contrasting the coefficient estimates across Panel A and Panel B in the first column of Table E1, relative subfield growth appears to be driven by stars whose death was anticipated. The effect in the case of sudden death is small in magnitude and imprecisely estimated.

[xii] Note that the variables in Table E1a pertain to subfield entry by non-collaborators only, except the first three, which correspond to the outcome variables in the right-most three columns of Table 3 (number of NIH grants acknowledged by articles in the subfield, in total, by collaborators only, and by non-collaborators only).

[xiii] On Figures E1, Panels A, B, and C, the peak appears roughly two to three years before the death, and not in the year of death. But recall that the source articles that generate the subfields in our data appeared in the window $[t_{yr\_death-5}; t_{yr\_death-1}]$. As such, the peak observed in these figures is an average of the peaks for subfields associated with sources published in the years $t_{yr\_death-5}, t_{yr\_death-4}, \ldots, t_{yr\_death-1}$.

As in Table 4, we parse every related article in the subfield to assign them into one of six mutually exclusive bins, based on their vintage-specific long-run citation impact: articles that fall in the bottom quartile of the citation distribution; in the second quartile; in the third quartile; articles that fall above the 75^{th} percentile, but below the 95^{th} percentile; articles that fall above the 95^{th} percentile, but below the 99^{th} percentile; and articles that fall above the 99^{th} percentile of the citation distribution. Decomposing this effect across the quantile bins as above reveals that the differences between the cases of sudden and anticipated death can be accounted for by shifts in activity for low-impact contributions. In the right tail of the distribution, there is very little evidence that the manner of superstar death matters at all for the fate of their subfields. In both cases, non-collaborators increase their relative contribution sharply—on the order of 40%.

Figure E2 and E3 display event study-style graphs in the spirit of Figure 2, Panel C. When using all publications (regardless of impact) as the metric of activity in a subfield (Figure E2), we can see that the upward trend is more pronounced (as well as statistically significant) in the case of anticipated events. When using only "top publications" (specifically, those in the upper 5 percentiles of the citation distribution, adjusted for each year of publication), the differences are less stark. Consistent with a dearth of statistical power, our ability to estimate these effects precisely is also limited. This convergence of the effect of death when focused on the upper tail of the impact distribution legitimates our choice to pool the data for sudden and anticipated events.

Consolidating vs. disruptive entry. The findings above do not imply that the published results of entrants necessarily contradict or overturn the prevailing scientific understanding and assumptions within a subfield. Direct evidence of these contributions' disruptive impact is elusive. To provide indirect evidence, we use the "disruptiveness" index (hereafter denoted d) recently proposed by Funk and Owen-Smith (2017), which seeks to capture whether an idea consolidates or destabilizes the status quo. d measures the extent to which the future ideas that build on the focal idea also rely on its acknowledged predecessors. In practice, for article i, it is defined as:

$$d_i = \frac{1}{n_i}\sum_{j=1}^{n}[1(j \vee K_i) - 1(j \wedge K_i)]$$

where j indexes the forward citing articles ($j = 1, \dots, n$), K_i is the set of articles $\{k_1, k_2, \dots, k_p\}$ that are (backwards) referenced within i, n_i is the number of forward citations to article i, $1(j \vee K_i)$ is equal to one if forward citing article j does not reference any of the articles in K_i, and $1(j \wedge K_i)$ is equal to one if forward citing article j does reference at least one of the articles in K_i. $d = 1$ for articles that are "maximally destabilizing," in the sense that there is no overlap between the articles referenced by the focal article and the references listed in the papers that cite it. In contrast, $d = -1$ for articles that are "maximally consolidating," in the sense that every citing article and the source have at least one reference in common.

We compute the d index for all related articles in our data (mean$= -.39$, median$= -.49$, s.d.$= .47$). We count the number of related articles that belong to a particular quantile bin of d. We create six non-overlapping bins: below the 10^{th} percentile of d, between the 10^{th} and the 25^{th} percentile, between the 25^{th} and the 50^{th} percentile, between the 50^{th} and the 75^{th} percentile, between the 75^{th} and the 95^{th} percentile, and above the 95^{th} percentile of d. In a final step, we roll up the outcome at the subfield-year level. We then run a separate regression with each of these six outcome measures, using the research design outlined in Section 4.1. As can be observed in Table E4, the relationship between star death and subfield entry by non-collaborators is non-monotonic in the extent to which it entails disrupting the paradigms of the treated subfields. The relationship is strongest for related articles that fall in the intermediate range of the "disruptiveness" metric. In contrast, the effect is zero and noisy when focusing on entry by both the most disruptive and the most consolidating articles.

Taken together, the results in Tables 5 and E4 paint a nuanced picture of directional change in the wake of superstar death. The new contributions do not represent a departure from the subfield's concerns. At the same time, the citation evidence makes it clear that these additional contributions often draw from more recent and different sources of knowledge for inspiration. Moreover, rather than to view these contributions as the expression of a Kuhnian paradigm shift within the subfield, it seems more appropriate to interpret them as reflecting the impact of a myriad "small r," permanent revolutions whereby new ideas come to the fore without necessarily eclipsing prior approaches.

Subfield characteristics. Table E5 examines how three different characteristics of subfields influence the magnitude of the treatment effect. We first inquire whether post-death entry by non-collaborators is more pronounced is subfields with forward momentum, relative to those where activity is relatively more subdued in the years leading up to the star's death. To create a metric of subfield "hotness," we compute the fraction of all papers in the subfield that were published in the window of five years before the star's death (or counterfactual death for the control subfields).[xiv] We then contrast the magnitude of the treatment effect in the subsamples of "hot" and "cold" subfields, respectively, by splitting the data across the median of the hotness covariate. Interestingly, the subfields with relatively less intense activity are driving the post-death entry effect. The treatment effect for hot subfields is half as small in magnitude, relative to that for cold subfields, and not statistically significant.

Next, we focus on the number of scientists trained by the star that had been active in the subfield before his death. We conjecture that the subfields of stars who produced many intellectual "offsprings" may be less welcoming to outsiders than those in which the stars did not train many graduate students or postdoctoral fellows. Of course, we do not have evidence that these individuals, once trained, remained intellectually beholden to the star. To identify trainees, we focus on the subset of collaborators who occupy the first author position in articles where the star occupies the last position; with the added stipulation that the coauthored publication appears in a window of ± three years around the year in which the collaborator's highest degree was received. We then count of the number of investigators trained by the star before his (possibly counterfactual) death. The results in Table E5 indicate that subfields that are relatively more endogamous (more than two trainees, the median of this covariate) experience elevated rates of entry after the star's death, relative to before. However, the difference between the coefficients corresponding to subfield with an above median of number of trainees versus below median number of trainees is not itself statistically significant.

Finally, we examine whether a star's level of *commitment* to a subfield moderates the extent of the post-death entry boost. Recall from Table 6 that the subfields where stars are relatively more *important* experience more entry following the star's death. A star could be important to a subfield, while not being fully committed to it, in the sense that his presence in the subfield represents only a small part of his overall published output. Empirically, we compute commitment as the fraction of a star's publications that fall into the focal subfield, and we split the data according to the median of this measure (which is equal to 0.14 in the data). The magnitudes of the treatment effects are very similar. What appears to be associated with the post-death entry boost is the star's importance to the subfield while alive, and not the extent of his commitment to it.

Impact of research infrastructure needs. Our analysis is limited to the life sciences and biomedical research. Though this area accounts for a large fraction of publicly funded, civilian research funding in the United States, it is not necessarily representative of all fields of science. In particular, some domains of research, like high-energy particle physics for example, require access to expensive and lumpy capital equipment, such as the Large Hadron Collider that came on line in 2009 at the cost $8 billion dollars (Stephan 2012). In contrast to the "big science," hypercollaborative projects that are emerging as the norm in these fields (e.g., Aad et al. 2015), academic life scientists require funding in sizable, but more modest amounts to do frontier research. In scientific domains where capital needs are lumpy, the phenomenon of entry in the wake of the passing of an eminent scientists may play out very differently, depending on the institutions that govern access to the scarce capital equipment.

[xiv] Only the articles in the subfield that were published before the death are taking into account when computing this ratio. The mean hotness across subfields is 0.61 (very similar to the median), with a standard deviation equal to 0.21.

Within biomedical research, large-scale clinical trials most closely resemble the characteristics of those other capital-intensive science fields. These necessitate a large infrastructure of data collection, monitoring, and management, which is why these activities are often consolidated in large cooperative groups such as the AIDS Clinical Trials Group, the Children's Oncology Group, or the Framingham Heart Study. *PubMed* has a "publication type" field which allows us to identify the subfields that are clinical-trial intensive (10% of the subfields) versus those that are not (the remaining 90%).

Table E6 replicates the results of Table 3 separately for these two subsamples. Unsurprisingly, our ability to estimate statistically significant effects is limited to the much larger set of non clinical trial-intensive subfields. That said, the magnitudes for the clinical trial-intensive subfields are very similar.

Star characteristics. We saw in Table 6 that the passing of stars that shone brighter while they were alive (measured by citations, publications or funding at death) appear to be driving much of the effect on non-collaborator entry. Tables E7 and E8 focus on other star characteristics that might moderate the core finding. The first two columns of Table E7 show that the subfields of relatively younger stars (those aged 60 and below at the time of their deaths, the median in our sample) account for much of the overall impact of death—the magnitude of the effect for older stars is very small and imprecisely estimated. However, there is potentially a distinction between being "young in the field" and simply being young. We measure experience in a subfield by capturing the year in which the star first published within it. Subfield experience varies from 1 to 38 years, with a median of seven and a mean of 8.36. The last two columns of Table E7 imply that the stars who are above median in subfield experience are associated with slightly more post-death entry, but the difference is very slight.

Table E8 brings more nuance to the analysis by focusing on the extent to which the star was leading vs. lagging the frontier of his subfields at the time of death. We develop two alternative measures of "distance to the frontier." We assume that frontier work will be more likely to reference more recent science, and alternatively will tend to be tagged by MeSH keyword combinations that are of more recent vintage. In a window of five years before the death, we then contrast the difference in reference vintage (respectively MeSH term combination vintage) for articles written by the star vs. articles written by all other authors. We then split subfields according to the median of this difference. Across all measures, the results in Table E8 tend to show that the effect of post-death entry are larger for those subfields where the star was leading when he passed, relative to those where his lead may have been slight or his research even staler than that of other researchers in the subfield.

Outsiders vs. competitors: A reprise. Recall that Figure 3 focused on the extent to which related authors were outsiders vs. previous incumbents in the subfields that expand in the wake of a star's death. For every related article, we matched their authorship roster to the Faculty Roster of the AAMC. Using the matched authors' past publication record, we can then ascertain the fraction of each related author's output that fall in the focal subfield. We then sorted each related article into 11 mutually exclusive bins: zero overlap (which corresponds to the bottom two quartiles of the overlap distribution), and a separate bin for every five percentiles above the median (50^{th} to 55^{th} percentile, 55^{th} to 60^{th} percentile,…, 95^{th} to 99^{th} percentile), as well as a top percentile bin. We then computed the corresponding measures of subfield activity by aggregating the data up to the subfield/year level. We presented the results graphically in Figure 3, Panel B, where each dot corresponds to the magnitude of the treatment effect in a separate regression with the outcome variable being the number of articles in each subfield that belong to the corresponding bins.

In Table E9, we provide, in regression table form this time, a variant of Table 3 where overlap is measured not just with respect to the focal subfield, but rather with respect to the combined subfields of a given star. We also simplify the number of bins, with only five: related articles by new scientists, related articles by scientists with zero overlap who have published in the past in other subfields, related articles by scientists in the third quartile of overlap, related articles by scientists whose past publication record puts them between the 75^{th} and 95^{th} percentile of the overlap distribution, and finally related articles by scientists whose past publication record puts them above the 95^{th} percentile of the overlap distribution. With this "global" measure of overlap, one can observe that the post-death entry boost is driven by scientists with no, or only limited past participation in the subfields where the star was active.

The lifecycle of stardom. The results in our manuscript naturally raise implications for welfare. We expound the view that once securely ensconced at the helm of their field, stars leverage their power for longer than a benevolent social planner might prefer. This argument would be less tenable if stars were able to remain at the peak of their intellectual abilities until the very twilight of their careers. To shed light on the career life cycle for superstars, we focus on the 5,878 control stars in our analytic sample, and construct a panel dataset of publications at the star scientist-year level.[xv] Using Poisson specifications, we then regress publication output onto year effects, indicator variables for degree (MD, PhD, MD/PhD), an indicator variable for female scientists, indicator variable for departmental affiliation (medicine vs. surgery vs. cell biology, etc.), indicator variables for the year in which the highest degree was received as well as 52 indicator variables for age effects (from age 29 to age 80, with ages below 29 absorbed in the omitted category).

Panel A of Figure E4 displays the estimates corresponding to the age effects when then outcome in the specification is the overall number of publications in a given year. Panel B restricts the outcome measure to publications whose number of long-run citations lies above the 95^{th} percentile of the vintage-specific citation distribution at the article level. Panel C proceeds in the same spirit, but focuses on even more impactful publications, those whose number of long-run citations lie above the 99^{th} percentile of the vintage-specific citation distribution at the article level. As can be observed in all three panels, the productive life cycle of stars follows an inverted U-shaped pattern, with a peak occurring earlier for highly cited publications, followed by a steeper drop off.

[xv] We eliminate the 452 extinct stars from the sample since their life cycle was interrupted prematurely.

Table E1a: Extended Descriptive Statistics, Subfield-level outcome variables

	Control Subfields			Treated Subfields		
	Mean	Median	Std. Dev.	Mean	Median	Std. Dev.
Baseline stock of related NIH grants, total	23.824	17	25.570	22.449	17	23.566
Baseline stock of related NIH grants, collaborators	4.876	2	6.952	4.446	2	6.011
Baseline stock of related NIH grants, non-collaborators	19.301	13	22.170	18.306	13	20.659
Baseline stock of related articles, bottom quartile of citation impact	6.614	4	8.322	6.741	4	8.611
Baseline stock of related articles, 2^{nd} quartile of citation impact	13.423	9	13.983	13.356	9	14.057
Baseline stock of related articles, 3^{rd} quartile of citation impact	20.100	14	19.051	19.996	14	18.937
Baseline stock of related articles, 75^{th} < citation impact< 95^{th} pctl.	21.762	16	19.810	21.271	16	19.289
Baseline stock of related articles, 95^{th} < citation impact< 99^{th} pctl.	5.233	3	5.933	5.108	3	5.844
Baseline stock of related articles, citation impact > 99^{th} pctl.	1.257	1	2.129	1.280	0	2.360
Baseline stock of related articles, outsiders	25.167	19	21.966	23.046	17	21.194
Baseline stock of related articles, incumbents	16.000	9	19.960	17.056	11	19.908
Baseline stock of related articles, proximate to source (cardinal measure)	31.353	24	31.179	32.022	25	31.854
Baseline stock of related articles, distant from source (cardinal measure)	37.037	19	49.598	35.730	18	48.119
Baseline stock of related articles, proximate to source (ordinal measure)	32.730	31	17.223	32.786	31	17.000
Baseline stock of related articles, distant from source (ordinal measure)	35.661	15	51.796	34.966	14	51.735
Baseline stock of related articles, references within subfield	54.627	42	48.492	53.963	41	47.581
Baseline stock of related articles, references outside subfield	13.764	7	19.080	13.789	7	19.159
Baseline stock of related articles, cites the star	32.332	22	34.199	31.076	22	32.141
Baseline stock of related articles, does not cite the star	36.058	24	37.875	36.677	24	39.633
Baseline stock of related articles, recent references	25.390	16	29.948	25.300	16	29.643
Baseline stock of related articles, old references	43.000	32	39.830	42.453	32	39.545
Baseline stock of related articles, recent MeSH terms (individual)	47.225	34	44.565	46.781	34	43.835
Baseline stock of related articles, old MeSH terms (individual)	20.465	7	32.090	20.318	7	30.955
Baseline stock of related articles, recent MeSH terms (combinations)	34.242	23	36.401	33.941	23	35.964
Baseline stock of related articles, old MeSH terms (combinations)	30.569	20	34.234	30.179	20	33.176
Baseline stock of related articles, with no star author	50.222	36	45.708	50.241	36	46.113
Baseline stock of related articles, with at least one star author	18.168	12	19.281	17.512	12	18.411
Baseline stock of related articles, with current elite author	62.275	46	55.514	61.728	45	55.169
Baseline stock of related articles, with no current or future elite author	2.699	1	4.855	2.657	1	4.715
Baseline stock of related articles, with future elite author	3.416	2	5.079	3.367	2	4.916

Note: All variables are limited to subfield activity by non-collaborators, unless otherwise specified.

Table E1b: Extended descriptive statistics, key covariates

	Control Subfields			Treated Subfields		
	Mean	Median	Std. Dev.	Mean	Median	Std. Dev.
Star-level						
Age at Death	58.100	58	8.795	58.100	58	8.796
Investigator Cumulative Nb. of Publications	164	131	123	170	143	118
Investigator Cumulative Nb. of Citations	12,141	8,010	12,938	11,580	8,726	10,212
Investigator Cumulative NIH Funding at Baseline	$18,784,517	$11,904,846	$25,160,518	$17,637,726	$12,049,690	$24,873,018
Star's number of past trainees (overall)	8.665	6	8.991	8.379	7	7.661
Subfield-level						
Importance of the star to the subfield	0.152	0	0.134	0.151	0	0.132
Commitment of the star to the subfield	0.160	0	0.149	0.157	0	0.149
Subfield coherence [PMRA-based measure]	0.602	1	0.131	0.603	1	0.128
Subfield coherence [citation-based measure]	-0.003	0	0.019	-0.003	0	0.023
Subfield cliquishness [Clustering Coefficient]	0.774	1	0.140	0.774	1	0.137
Cumulative Nb. of editorials by coauthors	122.453	35	217.358	118.844	39	201.803
Nb. of coauthors in study sections	0.324	0	0.846	0.369	0	0.971
% of subfield NIH funding controlled by the star's collaborators	0.285	0	0.315	0.269	0	0.307
Subfield "hotness"	0.597	1	0.212	0.596	1	0.217
Star's number of past trainees in the subfield	1.917	1	2.450	1.803	1	2.171
Years of experience in the subfield	8.277	7	5.750	8.493	7	6.078
Relative lead of the star in subfield [Individual MeSH measure]	0.045	-0	1.879	0.036	-0	1.741
Relative lead of the star in subfield [2-way combo MeSH measure]	-0.028	0	4.447	-0.089	0	4.334
Relative lead of the star in subfield [backward reference measure]	0.053	-0	6.902	0.227	-0	6.833

Note: This table reports summary statistics for all of the key covariates that we interact with the treatment effect in order to explore the underlying mechanisms of star death.

Table E2a: Correlation matrix, Subfield-level covariates

		(1)	(2)	(3)	(4)	(5)	(6)	(7)
(1)	Importance of the star to the subfield	1.00						
(2)	Commitment of the star to the subfield	0.34**	1.00					
(3)	Star's number of past trainees in the subfield	0.23**	0.24**	1.00				
(4)	Subfield coherence [PMRA-based measure]	-0.34**	-0.07**	0.04**	1.00			
(5)	Subfield coherence [citation-based measure]	-0.05**	-0.03**	-0.01	0.00	1.00		
(6)	Subfield cliquishness [clustering coefficient]	-0.25**	-0.37**	-0.40**	-0.10**	0.12**	1.00	
(7)	Cumulative nb. of editorials by coauthors	-0.06**	-0.20**	0.01*	0.07**	-0.03**	-0.00	1.00
(8)	Nb. of coauthors in study sections	-0.02**	-0.11**	0.13**	0.05**	-0.02**	-0.04**	0.48**
(9)	% of subfield NIH funding controlled by the star's collaborators	0.36**	0.15**	0.24**	-0.12**	-0.09**	-0.25**	0.10**
(10)	Subfield "hotness"	-0.03**	-0.07**	-0.10**	-0.05**	-0.09**	0.24**	-0.04**
(11)	Years of experience in the subfield	0.12**	0.30**	0.38**	0.07**	0.02**	-0.44**	0.09**
(12)	Relative lead of the star in subfield [individual MeSH measure]	0.01*	0.01	-0.00	-0.04**	0.01	0.02**	-0.01**
(13)	Relative lead of the star in subfield [2-way combo MeSH measure]	-0.00	0.01	-0.00	0.01**	0.01	-0.00	-0.01*
(14)	Relative lead of the star in subfield [backward reference measure]	-0.08**	-0.00	-0.04**	0.02**	0.02**	0.04**	-0.04**

		(8)	(9)	(10)	(11)	(12)	(13)	(14)
(8)	Nb. of coauthors in study sections	1.00						
(9)	% of subfield NIH funding controlled by the star's collaborators	0.12**	1.00					
(10)	Subfield "hotness"	-0.02**	-0.05**	1.00				
(11)	Years of experience in the subfield	0.09**	0.22**	-0.49**	1.00			
(12)	Relative lead of the star in subfield [individual MeSH measure]	0.00	0.00	-0.01†	-0.01†	1.00		
(13)	Relative lead of the star in subfield [2-way combo MeSH measure]	-0.01**	0.00	-0.05**	0.03**	0.29**	1.00	
(14)	Relative lead of the star in subfield [backward reference measure]	-0.03**	-0.06**	-0.10**	0.02**	0.05**	0.09**	1.00

Table E2b: Correlation matrix, Star-level covariates

		(1)	(2)	(3)	(4)	(5)
(1)	Age at Death	1.00				
(2)	Investigator Cumulative Nb. of Publications	0.40**	1.00			
(3)	Investigator Cumulative Nb. of Citations	0.21**	0.74**	1.00		
(4)	Investigator Cumulative NIH Funding at Baseline	0.38**	0.45**	0.34**	1.00	
(5)	Star's number of past trainees (overall)	0.33**	0.54**	0.56**	0.36**	1.00

† $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table E3: Scientific Impact of Entry by Non-Collaborators

	All Pubs	**Bttm. Quartile**	**2^{nd} Quartile**	**3^{rd} Quartile**	**Btw.75^{th} & 95^{th} pctl.**	**Btw. 95^{th} & 99^{th} pctl.**	**Above 99^{th} pctl.**
Panel A: Anticipated Death Events							
After Death	0.128^{**}	0.043	0.082^{*}	0.093^{*}	0.151^{**}	0.214^{**}	0.333^{**}
	(0.038)	(0.045)	(0.041)	(0.041)	(0.048)	(0.069)	(0.115)
Nb. of Investigators	4,018	3,982	4,018	4,016	4,013	3,946	3,214
Nb. of Fields	15,084	14,885	15,082	15,082	15,076	14,623	9,586
Nb. of Field-Year Obs.	554,869	547,637	554,795	554,795	554,573	537,883	352,571
Log Likelihood	-1,234,030	-315,200	-504,577	-633,777	-643,787	-234,637	-67,585
Panel B: Sudden Death Events							
After Death	0.026	$-0.102^{\dagger}$	-0.069	-0.040	0.090	0.243^{**}	0.310^{**}
	(0.048)	(0.057)	(0.055)	(0.054)	(0.057)	(0.075)	(0.116)
Nb. of Investigators	4,656	4,615	4,656	4,655	4,656	4,592	3,777
Nb. of Fields	17,549	17,253	17,539	17,545	17,549	17,063	11,331
Nb. of Field-Year Obs.	645,771	634,958	645,407	645,623	645,771	627,898	417,017
Log Likelihood	-1,396,961	-338,628	-563,370	-726,799	-756,820	-285,678	-83,118

Note: Estimates stem from conditional (subfield) fixed effects Poisson specifications. Like in Table 4 in the manuscript, the dependent variable is the total number of publications by non-collaborators in a subfield in a particular year, where these publications fall in a particular quantile bin of the long-run, vintage-adjusted citation distribution for the universe of journal articles in *PubMed*. In Panel A, the sample is limited to 1,576 subfields associated with 229 stars whose death is anticipated (along with the corresponding control subfields); and in Panel B, the sample is limited to 1,342 subfields associated with 185 stars whose death is sudden and unexpected (along with the corresponding control subfields). All models incorporate a full suite of year effects and subfield age effects, as well as a term common to both treated and control subfields that switches from zero to one after the death of the star. Exponentiating the coefficients and differencing from one yield numbers interpretable as elasticities. For example, the estimates in the first column of Panel A, imply that treated subfields see an increase in the number of contributions by non-collaborators after the superstar passes away—a statistically significant $100\times(\exp[0.128]-1)=13.66\%$.

Robust standard errors in parentheses, clustered at the level of the star scientist.
$^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$.

Table E4: Disruptive vs. Consolidating Entry

	Below 10^{th} pctl.	**Btw. 10^{th} & 25^{th} pctl.**	**Btw. 25^{th} and 50^{th} pctl.**	**Btw. 50^{th} and 75^{th} pctl.**	**Btw. 75^{th} and 95^{th} pctl.**	**Above 95^{th} pctl.**
Disruption Index d	d=-1	-1< d <-.74	-.74< d <-.50	-.50< d <-.14	-.14< d <0.53	d >0.53
After Death	0.005	0.071	0.139^{***}	0.154^{***}	0.121^{***}	0.002
	(0.041)	(0.041)	(0.034)	(0.031)	(0.034)	(0.041)
Nb. of Investigators	6,189	6,184	6,247	6,254	6,253	6,077
Nb. of Fields	33,610	33,868	34,183	34,205	34,147	30,889
Nb. of Field-Year Obs.	1,237,024	1,246,410	1,257,883	1,258,695	1,256,557	1,136,914
Log Likelihood	-670,691	-837,488	-1,218,093	-1,268,501	-1,134,304	-448,029

Note: Estimates stem from conditional (subfield) fixed effects Poisson specifications. The dependent variable is the total number of publications by non-collaborators in a subfield in a particular year, where these publications fall within a particular quantile bin of the Funk & Owen-Smith (2017) disruptiveness index, denoted by d. All models incorporate a full suite of year effects and subfield age effects, as well as a term common to both treated and control subfields that switches from zero to one after the death of the star. Exponentiating the coefficients and differencing from one yield numbers interpretable as elasticities. Robust standard errors in parentheses, clustered at the level of the star scientist. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$.

Table E5: Post-death entry and subfield characteristics

Metric of field Momentum	**"Hotness"**		**Number of Trainees**		**Commitment to the Field**	
	Below Median	Above Median	Below Median	Above Median	Below Median	Above Median
After Death	0.130^{**}	0.066	0.100^{*}	$0.059^{\dagger}$	$0.059^{\dagger}$	0.069
	(0.028)	(0.044)	(0.041)	(0.035)	(0.032)	(0.046)
Nb. of Investigators	4,870	4,694	3,566	4,881	4,477	4,520
Nb. of Fields	17,427	16,791	8,652	25,566	17,072	17,146
Nb. of Field-Year Obs.	642,219	616,957	317,813	941,363	627,355	631,821
Log Likelihood	-1,453,789	-1,137,226	-677,372	-2,085,856	-1,345,958	-1,413,964

Note: Estimates stem from conditional (subfield) fixed effects Poisson specifications. The dependent variable is the total number of publications by non-collaborators within a subfield in a particular year. Each pair of columns splits the sample across the median of a particular covariate for the sample of subfields (treated and control) in the baseline year. The first set of two columns examines differences in the extent to which the "hotness" of the subfield—defined as the fraction of the subfield's activity that falls within the time window of five years before the star's death—influences the rate at which non-collaborators enter the field after the star passes away. The second set of columns examines the impact of having former trainees of the star in the subfield. The final set of columns splits the sample according to the degree of commitment of the star to the subfield (i.e., the fraction of his/her output that falls within the subfield. All models incorporate a full suite of year effects and subfield age effects, as well as a term common to both treated and control subfields that switches from zero to one after the death of the star. Exponentiating the coefficients and differencing from one yield numbers interpretable as elasticities. Robust standard errors in parentheses, clustered at the level of the star scientist. $^{\dagger}p<0.10$, $^{*}p<0.05$, $^{**}p<0.01$.

Table E6: Impact of Research Infrastructure Needs

	Clinical Trial-intensive			Other		
	All Authors	Collabs. Only	Non-Collabs. Only	All Authors	Collabs. Only	Non-Collabs. Only
After Death	0.061	-0.147	$0.086^{\dagger}$	$0.060^{\dagger}$	-0.262^{**}	0.095^{**}
	(0.051)	(0.102)	(0.052)	(0.031)	(0.065)	(0.032)
Nb. of Investigators	1,739	1,666	1,739	5,753	5,630	5,753
Nb. of Fields	3,437	3,309	3,437	30,781	29,787	30,781
Nb. of Field-Year Obs.	125,919	121,230	125,919	1,133,257	1,096,675	1,133,257
Log Likelihood	-315,048	-77,390	-302,267	-2,628,821	-660,968	-2,510,273

Note: Estimates stem from conditional (subfield) fixed effects Poisson specifications. The dependent variable is the total number of publications in a subfield in a particular year. The first set of three columns replicate our benchmark specifications (Table 3, columns 1, 2, and 3) on the sample of subfields where research often entails performing large scale clinical trials. The second set of three columns replicate the benchmark specifications on the sample of subfields where research seldom entails performing large-scale clinical trials. Clinical trial publications were identified using the publication type field in *PubMed*. All models incorporate a full suite of year effects and subfield age effects, as well as a term common to both treated and control subfields that switches from zero to one after the death of the star, to address the concern that age, year and individual fixed effects may not fully account for trends in subfield entry around the time of death for the deceased star. Robust standard errors in parentheses, clustered at the level of the star scientist. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$.

Table E7: Influence of star age and in-field experience

	Star Birth Age at Time of Death		Star Experience in the Field at Time of Death	
	Younger than 61	61 or Older	Recent (less than 7 years)	Established (more than 7 years)
After Death	0.108^{**}	0.009	$0.061^{\dagger}$	0.092^{*}
	(0.041)	(0.041)	(0.037)	(0.036)
Nb. of Investigators	5,542	1,936	5,166	4,257
Nb. of Fields	27,022	7,196	17,933	16,285
Nb. of Field-Year Obs.	995,153	264,023	659,252	599,924
Log Likelihood	-2,178,601	-581,832	-1,376,994	-1,348,968

Note: Estimates stem from conditional (subfield) fixed effects Poisson specifications. The dependent variable is the total number of publications by non-collaborators within a subfield in a particular year. All models incorporate a full suite of year effects and subfield age effects, as well as a term common to both treated and control subfields that switches from zero to one after the death of the star, to address the concern that age, year and individual fixed effects may not fully account for trends in subfield entry around the time of death for the deceased star. Exponentiating the coefficients and differencing from one yield numbers interpretable as elasticities. Robust standard errors in parentheses, clustered at the level of the star scientist. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$.

Table E8: Star's leadership relative to the frontier in his/her subfield

Metric of distance to the subfield frontier	Vintage of cited references		Vintage of MeSH terms			
			Individual		2-way combinations	
	Lagging	Leading	Lagging	Leading	Lagging	Leading
After Death	0.117**	0.154*	0.063	0.192**	0.094†	0.167**
	(0.037)	(0.072)	(0.047)	(0.049)	(0.057)	(0.041)
Nb. of Investigators	3,373	3,075	3,328	3,210	3,333	3,216
Nb. of Fields	9,226	7,664	8,647	8,243	8,762	8,128
Nb. of Field-Year Obs.	339,900	282,526	318,626	303,800	322,838	299,588
Log Likelihood	-775,180	-618,943	-713,539	-682,532	-729,341	-666,577

Note: Estimates stem from conditional (subfield) fixed effects Poisson specifications. The dependent variable is the total number of publications by non-collaborators within a subfield in a particular year. We develop two alternative measures of "distance to the frontier." We assume that frontier work will be more likely to reference more recent science, and alternatively will tend to be tagged by MeSH keyword combinations that are of more recent vintage. In a window of five years before the death, we then contrast the difference in reference vintage (respectively MeSH term combination vintage) for articles written by the star vs. articles written by all other authors. We then split subfields according to the median of this difference. All models incorporate a full suite of year effects and subfield age effects, as well as a term common to both treated and control subfields that switches from zero to one after the death of the star, to address the concern that age, year and individual fixed effects may not fully account for trends in subfield entry around the time of death for the deceased star. Robust standard errors in parentheses, clustered at the level of the star scientist. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$.

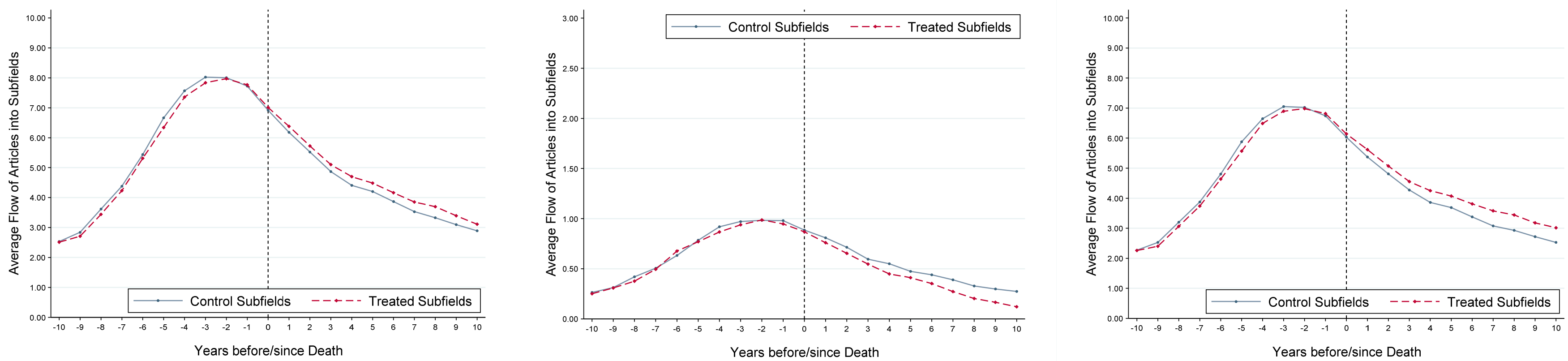
Table E9: Influence of field overlap between related authors and the stars on the rate of entry into subfields

	New Scientists	Below Median	Btw. 50^{th} and 75^{th} pctl.	Btw. 75^{th} and 95^{th} pctl.	Above 95^{th} pctl.
Intellectual Overlap x	Not Defined	x=0	0<x<6.35%	6.35%<x<36.70%	x>36.70%
After Death	0.081	0.113**	0.096*	-0.000	-0.075
	(0.082)	(0.028)	(0.038)	(0.061)	(0.128)
Nb. of Investigators	4,724	6,260	6,167	5,638	3,622
Nb. of Fields	16,961	34,216	33,688	29,845	15,241
Nb. of Field-Year Obs.	625,066	1,259,102	1,239,873	1,098,754	561,888
Log Likelihood	-88,890	-1,508,995	-970,344	-633,095	-149,524

Note: This table displays a variation of the results depicted in Figure 3, Panel B in regression form. Estimates stem from conditional (subfield) fixed effects Poisson specifications. The dependent variable is the total number of publications by non-collaborators within a subfield in a particular year, broken into five bins: publications by new scientists; publications that fall below the median of our measure of field overlap between the star and the related investigators identified on these articles' authorship roster; publications that fall in the third quartile of the field overlap measure; publications that fall in the fourth quartile but below the top ventile of the field overlap measure; and finally publications that fall in the top ventile of the measure. In contrast to Figure 3, in this case overlap has been defined with respect to the "global" subfield that encompasses all the subfields of a given star in the data, as opposed to the "local" measure where overlap with the focal subfield determines the extent of overlap. All models incorporate a full suite of year effects and subfield age effects, as well as a term common to both treated and control subfields that switches from zero to one after the death of the star, to address the concern that age, year and individual fixed effects may not fully account for trends in subfield entry around the time of death for the deceased star. Robust standard errors in parentheses, clustered at the level of the star scientist. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$.

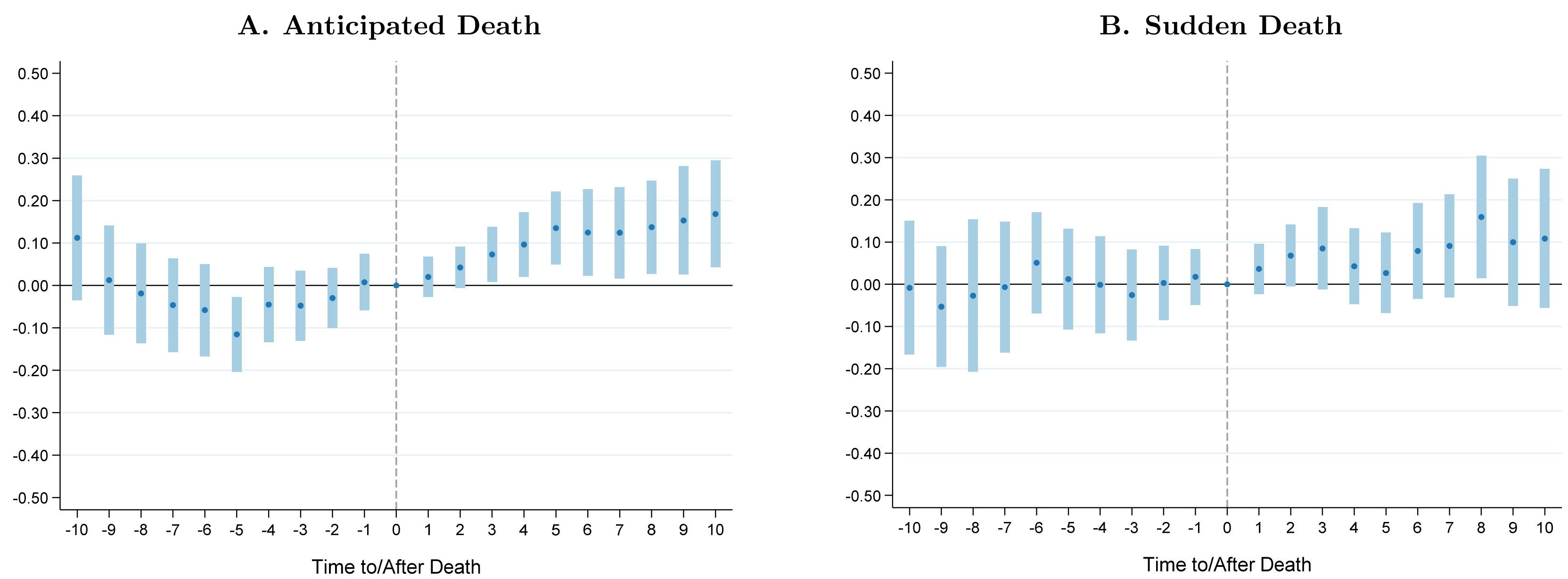
Figure E1
Subfield Growth and Decline—Raw Data

A. All Authors   B. Collaborators   C. Non-Collaborators



Note: Panels A, B, and C show the path of mean publication activity for treated and control subfields around the year of star death, broken down by total number of publications in the subfield (Panel A), number of publications in the subfield with a coauthor of the star (Panel B), and number of publications in the subfield without any coauthor of the star (Panel C).

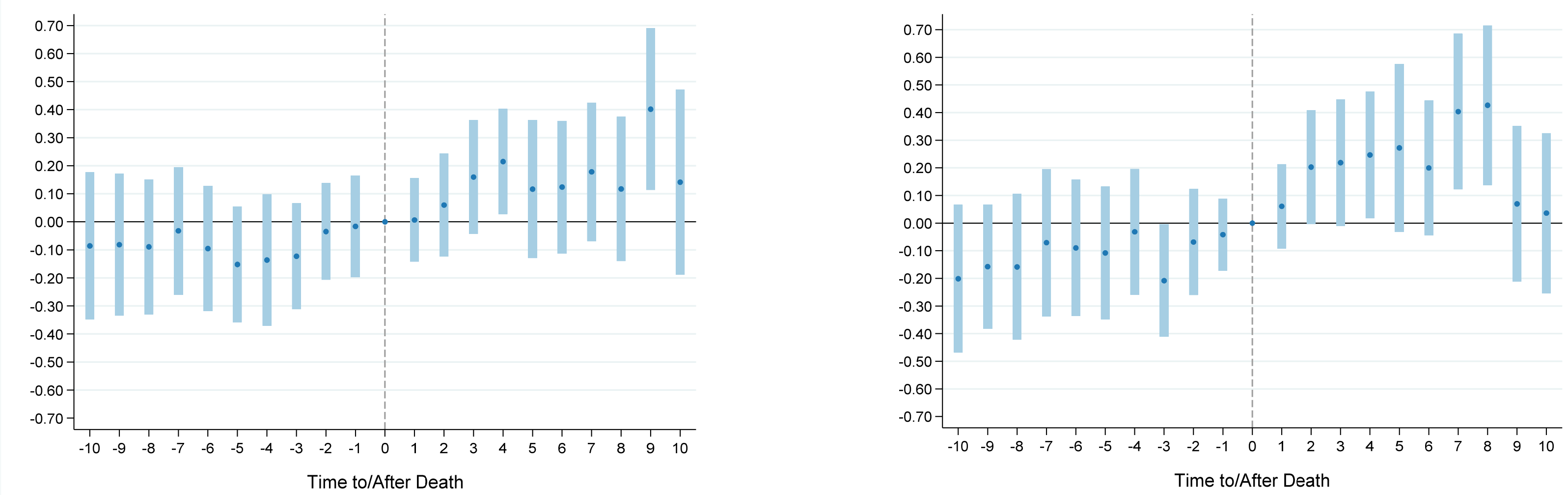
Figure E2: Effect of Star Scientist Death on Subfield Growth and Decline
Non-collaborator Activity Only—All Publications



Note: The graphs in this figure are patterned after Panel C in Figure 2 in the main body of the manuscript. The dark blue dots correspond to coefficient estimates stemming from conditional (subfield) fixed effects Poisson specifications in which publication flows by non-collaborators within a subfield are regressed onto year effects, subfield age effects, as well as 20 interaction terms between treatment status and the number of years before/after the death event (the indicator variable for treatment status interacted with the year of death is omitted). The specifications also include a full set of lead and lag terms common to both the treated and control subfields to fully account for transitory trends in subfield activity around the time of the death. These regressions are run separately on the subsample of subfields associated with stars whose death was anticipated (and their controls—Panel A), and on the subsample of subfields associated with stars whose death was sudden (and their controls—Panel B). The 95% confidence interval (corresponding to robust standard errors, clustered around star scientist) around these estimates is plotted with the vertical light blue lines.

Figure E3: Effect of Star Scientist Death on Subfield Growth and Decline
Non-collaborator Activity Only—Top 5% publications by citation

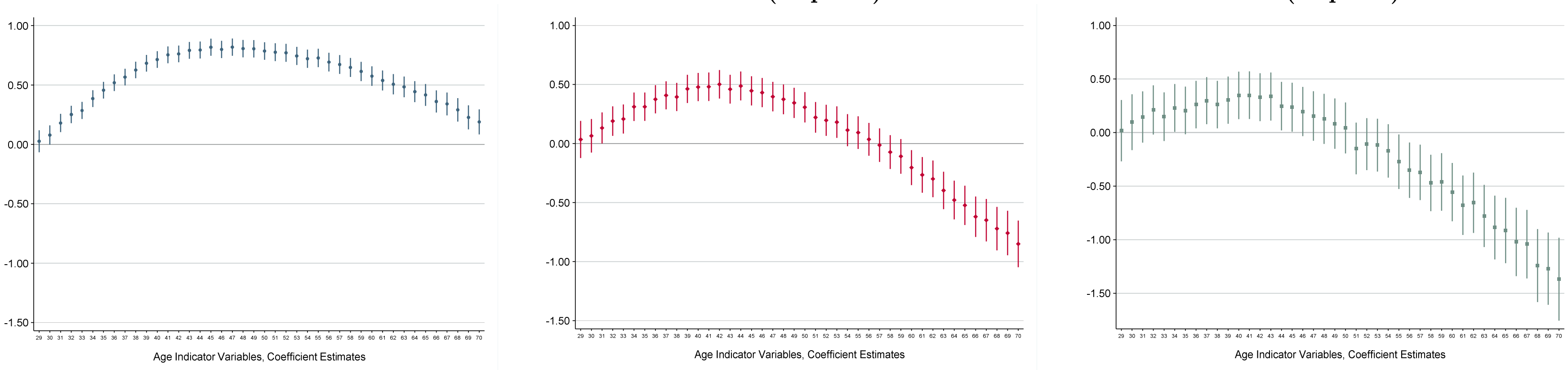
A. Anticipated Death **B. Sudden Death**



Note: The graphs in this figure are patterned after Panel C in Figure 2 in the main body of the manuscript. The dark blue dots correspond to coefficient estimates stemming from conditional (subfield) fixed effects Poisson specifications in which the flows of highly-cited publications (top 5% of the vintage-specific citation distribution) by non-collaborators within a subfield are regressed onto year effects, subfield age effects, as well as 20 interaction terms between treatment status and the number of years before/after the death event (the indicator variable for treatment status interacted with the year of death is omitted). The specifications also include a full set of lead and lag terms common to both the treated and control subfields to fully account for transitory trends in subfield activity around the time of the death. These regressions are run separately on the subsample of subfields associated with stars whose death was anticipated (and their controls—Panel A), and on the subsample of subfields associated with stars whose death was sudden (and their controls—Panel B). The 95% confidence interval (corresponding to robust standard errors, clustered around star scientist) around these estimates is plotted with the vertical light blue lines.

Figure E4
The Life Cycle of Stardom

A. All Publications | **B. Highly-cited Publications (Top 5%)** | **C. Highly-cited Publications (Top 1%)**



Note: For the sample of 5,878 control superstars, we create a panel dataset at the scientist-year level. We regress (i) publication output in a given year (Panel A) and (ii) highly-cited publications in a given year (Panels B and C) onto year effects, indicator variables for degree (MD, PhD, MD/PhD), an indicator variable for female scientists, indicator variable for departmental affiliation (medicine vs. surgery vs. cell biology, etc.), indicator variables for the year in which the highest degree was received as well as 52 indicator variables for age effects (from age 29 to age 80, with ages below 29 absorbed in the omitted category). In Panel B, a publication is deemed to be highly cited if it falls above the 95^{th} percentile of the vintage-specific citation distribution at the article level. In Panel C, a publication is deemed to be highly cited if it falls above the 99^{th} percentile of the vintage-specific citation distribution at the article level. The above plots display the estimates for the age indicator variables up to the age of 70 (to preserve the same scale across the three figures), together with their associated 95% confidence interval. The list of covariates is strictly identical across the three panels.

Appendix F: Robustness Checks

Balanced panel. With treatment events staggered over time, a concern with the dynamic specifications summarized in Figure 2 is that the magnitude of the treatment effect might not be stable over time. Because our observation period stops in 2006, the lead terms far away from death are identified by only a subsample of the data (see Figure F1). Could such heterogeneity confound the true dynamics, for example if deaths that occurred earlier in the sample have a bigger effect? To address this concern, we extend the observation period used to generate the event study graphs in Figure 2 from 2006 to 2012, resulting in a sample that is almost perfectly balanced in a window of ten years before to ten years after the death of a superstar. As can be seen in Figure F2, which replicates Figure 2 in all respects except the length of the analytic sample, the results change very little.

This figure begs another question: why not simply use this longer observation period as the default throughout the paper? There are two reasons. First, we cannot identify collaborator status reliably after 2006 because this is the last year of the data in our version of the AAMC Faculty Roster. Second, whereas we can account precisely for the employment status of the control superstars up to 2006 (the year during which we coded their CVs), some may retire, or even die in the years that follow, raising the specter that their subfields are not adequate controls during the 2007-2012 time period. As a result, we quickly revert back to the observation window 1965-2006 in all that follows.

Main results, rolled up to the scientist-level of analysis. The treatment variable exhibits variation at the level of the star scientist, and not at the level of the subfield-star pair. Of course, we cluster the standard errors at the star level, and we exploit the differential position of a star across his subfields to shed light on mechanisms. But do our main results survive when the data is "rolled up" to the star-year level of the analysis? To probe the robustness of our benchmark set of results, we lump all related articles for each star together as if they belonged to a single subfield. Nevertheless, the results in Table F1 and Figure F3 are very similar to those in Table 3 and Figure 2, both in terms of magnitude and statistical significance. One exception is the coefficient on the effect of entry by collaborators in Table F1, which is negative as expected, but smaller in magnitude, relative to the corresponding coefficient in Table 3.

Alternate functional forms. Despite its robustness and appropriateness for the analysis of skewed positive outcomes, the conditional fixed effects Poisson model of Hausman et al. (1984) has an important shortcoming: subfields for which there is no variation in the outcome during the observation period (for example, because the outcome is uniformly zero) drop out of the sample. This is why the number of observations in many tables varies slightly from column to column. Fixed-effects OLS models do not suffer from this limitation. In Table F2 and Figure F4, we examine the sensitivity of our benchmark set of results to the choice of alternative functional forms. In the three columns to the left, we simply use the "raw" number of articles in the subfield as the outcome, and perform estimation by OLS. Of course, the estimates are not directly interpretable in terms of elasticities. At the mean of the data, however, the treatment effect in the third column implies that subfield entry by non-collaborating authors expands by $0.409/3.335 = 12.26\%$, which is not all that different from the 8.2% reported in Table 3.

In the three columns to the right, Table F2 reports results corresponding to OLS estimation, but this time with the outcome variables transformed using the inverse hyperbolic sine function (Burbidge et al. 1988).[xvi] In this case, coefficient estimates can be interpreted as elasticities, as an approximation. They are quite similar once again to those reported in Table 3, except for the effect on entry by collaborators, which is smaller in magnitude.

[xvi] $\sinh^{-1}(x) = \ln(x + \sqrt{x^2+1})$. Unlike the log of x, the inverse hyperbolic sine is defined at zero, which is attractive here because a substantial proportion of the subfields in the data display no activity in a particular year. For example, all subfields obviously see entry over the entire observation period, and yet in 31.33% of the subfield-year observations, the number of articles entering is zero.

Figure F4 presents dynamic analogs of the results in the the third column (Panel A) and sixth column (Panel B) of Table F2. In the case of the raw outcomes (Panel A), one can detect a trend in outcome before the event, though it is not estimated precisely. The results using the inverse hyperbolic sine transformation (Panel B) exhibit no evidence of a pre-trend.

Size of the control group. The first three columns in Table F3 drop from the sample all the control subfields, but are otherwise analogous to the core results presented in Table 3, Panel A. In these specifications, subfields who were treated in the past or will be treated in the future serve as implicit controls for the subfields currently experiencing the death of their associated star. The results are qualitatively similar to those displayed in Table 3. However, the corresponding event study graphs (Figure F5) clearly show that dropping the control group from the estimation sample produces pre-event trends that cast doubt on a research design based on a single level of difference. This provides a clear rationale for our preferred research design, which adds an additional level of difference to the data—that provided by control subfields.

The second set of three columns in Table F3 attempt to replicate the results of Table 3 in a sample such that for each treated subfield, there is exactly one control subfield (selected at random from the set of control subfields for each treated source). The magnitudes are qualitatively similar to those observed in Table 3, but the standard errors are larger. We conclude that the approximate 1 : 10 ratio of treated to control subfields is important insofar as it provides the statistical power to estimate the post-death term that is common between treated and control subfields, and to do so net of the subfield age and calendar year dynamics.

Are death events exogenous? Could some of the deaths in our sample be caused by stress as others are seeking to break a stars' hold on a field? Chronic stress can lead to a wide range of adverse health conditions. Most of these conditions diminish quality of life but not mortality per se. The most notable link between stress and death is through heart disease. Thus one possibility is that stress increases the risk of a heart attack. 14% of the extinct superstars (who account for 16.75% of the treated subfields) die of a heart attack.

In Table F4 (three leftmost columns), we replicate the results of Table 3 while excluding these subfields. The point estimates are slightly larger in magnitude, and also slightly more precisely estimated when excluding the subfields associated with heart attack events. Of course, there may be other more indirect channels through which stress can precipitate death. From a study design perspective, we would be more concerned with this threat to identification if subfield growth was trending upward before the death. But from the event study-type figures we present (Figure 2, as well as numerous variations in Appendices E and F), this does not appear to be the case.

Multi-disciplinary source articles and the validity of the control group. Multi-disciplinary journals such as *PNAS*, *Science*, or *Nature* account for 10% of the subfields in our data.[xvii] This could be problematic insofar as these prestigious outlets publish articles in all scientific fields, and we recruit control source articles from the same journal and year as that of the treated source article. Take the source paper by Chu et al. (1998)—already used as an example in Appendix D—which appeared in the issue of *Science* dated October 23^{rd} of that year. The same issue includes a paper with the title *"Climate and groundwater recharge during the last glaciation in an ice-covered region"* and another called *"Self-organized growth of three- dimensional quantum-Dot crystals with fcc-like stacking and a tunable lattice constant."* It would not seem advisable to use one of these as the source for a control subfield, since they do not pertain to the life sciences, even under the most expansive definition of this term.

This is not an issue in practice, since to qualify as a control, it is not sufficient for a candidate source article to appear in the same journal and year as its treated counterpart. In addition, we impose the requirement that one of our 12,000+ still alive superstars is in last authorship position. This will filter out of the set of potential controls any non-biomedical articles that appear in these outlets since all the stars in our data

[xvii]Note that *PLoS One*, a very large multidisciplinary journal, does not contribute any source article in our sample. This is because it was founded in 2006, and the latest year of publication for one of the source articles (treated or control) is 2002 (one year before the latest year of death, which is 2003).

(deceased or not) are life scientists. We have also replicated the benchmark results excluding the subfields that are associated with a source article published in either *Science*, *Nature*, or *PNAS*. The results are displayed in the three rightmost columns in Table F4. The point estimates are very similar to those we obtain in our benchmark set of analyses (columns 1, 2, and 3 of Table 3).

Source articles, with and without abstracts. In Table F5, we perform one analysis that (imperfectly) tries to assess the sensitivity of our results to the use of author-chosen information to delineate the set of intellectual neighbors in a subfield. Ten percent of the subfields in the data radiate from source articles for which *PubMed* does not have an abstract. For these subfields, PMRA must therefore make do without abstract words (i.e., relying solely on title words and MeSH terms) to return a set of neighbors. We reproduce our benchmark set of specifications (the first panel of Table 3) on the set of subfields radiating from source articles with and without abstract information. As can be seen above, the estimates for the sample restricted to abstractless source articles are less precisely estimated than for the sample restricted to the much larger number of subfields associated with source articles that have an abstract. The magnitudes in both cases, however, are quite similar, which we find reassuring.

Table 7 estimated on the subsample of less-well cited stars. Table 7 provides evidence that subfield entry is more pronounced after the death of an eminent scientist when the subfield can be perceived as less coherent, or when the colleagues of the star are less able to exert control over critical resources after he has passed away. However, the sample for these results was limited to the subfields of well-cited stars (those above the median by cumulative citations in the sample, in the year of death). For completeness, Table F6 provides an exact analog to Table 7, except that in this case the sample is limited to the subfields of less well-cited stars (those below the median by cumulative citations in the sample, in the year of death).

The results in this subsample are less consistent across measures than was the case for the more eminent stars. Many pairs of columns do not show notable differences between more coherent and less coherent subfields, or between more indirectly controlled vs. less indirectly controlled subfields. In two instances, however, the direction of the results is opposite to that observed in Table 7. First, subfields that were relatively less consolidated according to the metric of Funk and Owen-Smith (2017) see increased entry after the passing of a less eminent star (second and third columns of Panel A). Second, subfields in which the less eminent star had important coauthors sitting on NIH study sections in the last five years of his life also experience elevated rated of entry post-death (second and third columns of Panel B).

Table F1: Impacts at the level of the star scientist

	Publication Flows			NIH Funding Flows (Nb. of Awards)		
	All Authors	Collabs. Only	Non-Collabs. Only	All Authors	Collabs. Only	Non-Collabs Only
	(1)	(2)	(3)	(4)	(5)	(6)
After Death	0.227**	-0.121	0.249**	0.248**	-0.092	0.272**
	(0.056)	(0.088)	(0.055)	(0.059)	(0.098)	(0.058)
Nb. of Stars	6,369	6,369	6,369	5,440	5,172	5,427
Nb. of Star-Year Obs.	801,654	801,654	801,654	15,469	14,589	15,436
Log Likelihood	-2,444,982	-663,888	-2,262,127	479,539	452,259	478,516

Note: Estimates stem from conditional (star) fixed effects Poisson specifications. The dependent variable is the total number of publications in the collection of subfields in which the star (deceased or not) was active in a particular year. All models incorporate a full suite of year effects and star career age effects, as well as term common to both treated and control stars that switches from zero to one after the (possibly counterfactual) death of the star. Exponentiating the coefficients and differencing from one yield numbers interpretable as elasticities. For example, the estimates in column (3) imply that treated stars see an increase in the number of contributions by non-collaborators in their fields—a statistically significant 100×(exp[0.249]-1)=28.27%. Robust standard errors in parentheses, clustered at the level of the star scientist. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$.

Table F2: Alternate Functional Forms

	OLS (in levels)			OLS (inverse hyperbolic sine)		
	All Authors	Collabs. Only	Non-Collabs. Only	All Authors	Collabs. Only	Non-Collabs. Only
After Death	0.334**	-0.145**	0.409**	0.032	-0.054**	0.065**
	(0.108)	(0.032)	(0.100)	(0.025)	(0.014)	(0.024)
Nb. of Investigators	6,260	6,260	6,260	6,260	6,260	6,260
Nb. of Fields	34,218	34,218	34,218	34,218	34,218	34,218
Nb. of Field-Year Obs.	1,259,176	1,259,176	1,259,176	1,259,176	1,259,176	1,259,176
Mean of the Depndt. Var.	3.757	0.606	3.335	1.407	0.289	1.315
Adjusted R^2	0.428	0.380	0.400	0.555	0.329	0.523

Note: Estimates stem from (subfield) fixed effects OLS specifications. In columns 1, 2, and 3, the dependent variable is the number of publications in a subfield in a particular year. In columns 4, 5, and 6, the dependent variable is the inverse hyperbolic sine of the number of publications in a subfield in a particular year. All models incorporate a full suite of year effects and subfield age effects, as well as a term common to both treated and control subfields that switches from zero to one after the death of the star, to address the concern that age, year and individual fixed effects may not fully account for trends in subfield entry around the time of death for the deceased star. Robust standard errors in parentheses, clustered at the level of the star scientist. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$.

Table F3: Alternate Control Groups

	No Controls			1:1 Ratio Treated to Control Subfields		
	All Authors	Collabs. Only	Non-Collabs. Only	All Authors	Collabs. Only	Non-Collabs. Only
After Death	0.052	-0.312**	0.058†	0.023	-0.205**	0.049
	(0.033)	(0.045)	(0.034)	(0.033)	(0.061)	(0.034)
Nb. of Investigators	452	430	452	2,557	2,439	2,557
Nb. of Fields	3,076	2,885	3,076	6,152	5,800	6,152
Nb. of Field-Year Obs.	111,708	104,705	111,708	223,416	210,502	223,416
Log Likelihood	-255,523	-57,768	-245,596	-520,195	-118,841	-498,256

Note: Estimates stem from conditional (subfield) fixed effects Poisson specifications. The dependent variable is the total number of publications in a subfield in a particular year. All models incorporate a full suite of year effects and subfield age effects. Columns 4, 5, and 6 also include a term common to both treated and control subfields that switches from zero to one after the death of the star, to address the concern that age, year and individual fixed effects may not fully account for trends in subfield entry around the time of death for the deceased star. Robust standard errors in parentheses, clustered at the level of the star scientist. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$.

Table F4: Additional Robustness Checks

	Excluding Heart Attacks			Excluding Multi-disciplinary Journals		
	All Authors	Collabs. Only	Non-Collabs. Only	All Authors	Collabs. Only	Non-Collabs. Only
After Death	0.060*	-0.235**	0.093**	0.074*	-0.212**	0.105**
	(0.030)	(0.063)	(0.030)	(0.029)	(0.058)	(0.030)
Nb. of Investigators	5,817	5,685	5,817	5,811	5,670	5,811
Nb. of Fields	26,728	25,793	26,728	28,707	27,741	28,707
Nb. of Field-Year Obs.	983,372	948,973	983,372	1,056,127	1,020,609	1,056,127
Log Likelihood	-2,243,461	-562,978	-2,147,307	-2,455,832	-616,652	-2,355,142

Note: Estimates stem from conditional (subfield) fixed effects Poisson specifications. The dependent variable is the total number of publications by non-collaborators in a subfield in a particular year. All models incorporate a full suite of year effects and subfield age effects, as well as a term common to both treated and control subfields that switches from zero to one after the death of the star, to address the concern that age, year and individual fixed effects may not fully account for trends in subfield entry around the time of death for the deceased star. Robust standard errors in parentheses, clustered at the level of the star scientist. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$.

Table F5: Additional Robustness Checks (cont'd)

	Only source with abstracts			**Only source without abstracts**		
	All Authors	Collabs. Only	Non-Collabs. Only	All Authors	Collabs. Only	Non-Collabs. Only
After Death	0.055*	-0.234**	0.089**	0.129	-0.224†	0.148†
	(0.028)	(0.061)	(0.028)	(0.081)	(0.118)	(0.083)
Nb. of Investigators	6,009	5,905	6,009	1,549	1,399	1,549
Nb. of Fields	30,787	30,052	30,787	3,431	3,044	3,431
Nb. of Field-Year Obs.	1,132,555	1,105,538	1,132,555	126,621	112,367	126,621
Log Likelihood	-2,621,169	-689,447	-2,502,613	-276,654	-46,146	-266,293

Note: Estimates stem from conditional (subfield) fixed effects Poisson specifications. The dependent variable is the total number of publications by non-collaborators in a subfield in a particular year. All models incorporate a full suite of year effects and subfield age effects, as well as a term common to both treated and control subfields that switches from zero to one after the death of the star, to address the concern that age, year and individual fixed effects may not fully account for trends in subfield entry around the time of death for the deceased star. Robust standard errors in parentheses, clustered at the level of the star scientist. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$.

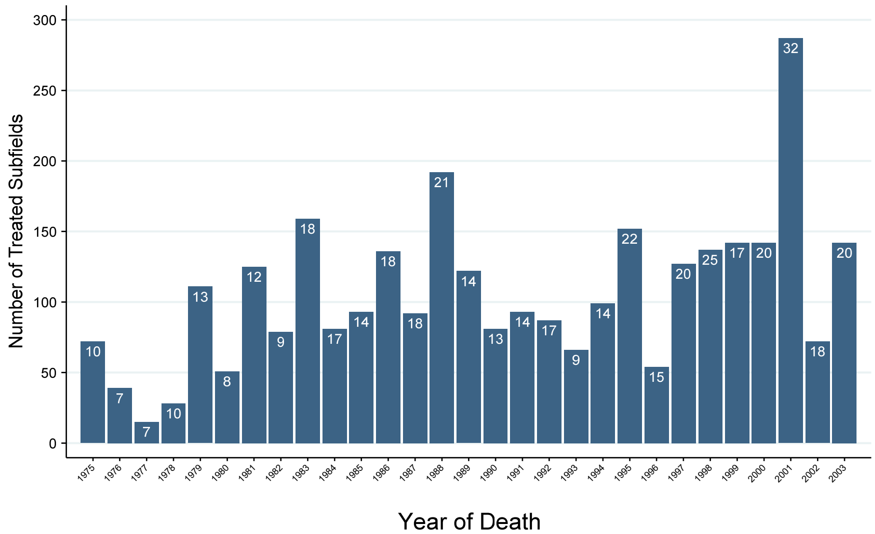
Table F6: The Nature of Entry Barriers for Less Cited Stars

Panel A	Subfield Coherence					
	PMRA-based definition		Citation-based definition		Cliquishness	
	Below Median	Above Median	Below Median	Above Median	Below Median	Above Median
After Death	0.044	0.024	-0.021	0.128**	-0.018	0.052
	(0.052)	(0.047)	(0.047)	(0.045)	(0.053)	(0.040)
Nb. of Investigators	2,131	2,257	2,118	2,232	2,087	2,263
Nb. of Fields	8,068	9,260	8,191	9,137	9,181	8,147
Nb. of Field-Year Obs.	296,675	340,075	301,130	335,620	337,770	298,980
Log Likelihood	-604,994	-746,571	-690,078	-673,587	-749,640	-595,838

Panel B	Indirect Control through Collaborators					
	Editorial Channel		NIH Study Section Channel		Fraction of Subfield NIH Funding	
	Below Median	Above Median	Below Median	Above Median	Below Median	Above Median
After Death	-0.041	0.072	-0.003	0.149†	0.029	0.055
	(0.063)	(0.052)	(0.050)	(0.083)	(0.049)	(0.059)
Nb. of Investigators	1,024	2,455	2,279	1,367	1,997	2,135
Nb. of Fields	5,719	11,609	12,153	5,175	7,806	9,522
Nb. of Field-Year Obs.	210,920	425,830	446,939	189,811	287,089	349,661
Log Likelihood	-495,980	-892,355	-1,000,972	-393,326	-646,673	-713,420

Note: Estimates stem from conditional (subfield) fixed effects Poisson specifications. The dependent variable is the total number of publications by non-collaborators in a subfield in a particular year. The sample is limited to the subfields in which the least eminent among the stars were active (specifically, below the median of the "cumulative citations up to the year of death" metric). Each pair of columns splits the sample across the median of a particular covariate for the sample of fields (treated and control) in the baseline year. For example, the first two columns of Panel B compare the magnitude of the treatment effect for stars whose collaborators have written an above-median number of editorials in the five years preceding the superstar's death, vs. a below-median number of editorials. All models incorporate a full suite of year effects and subfield age effects, as well as a term common to both treated and control subfields that switches from zero to one after the death of the star. Exponentiating the coefficients and differencing from one yield numbers interpretable as elasticities. Robust standard errors in parentheses, clustered at the level of the star scientist. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$.

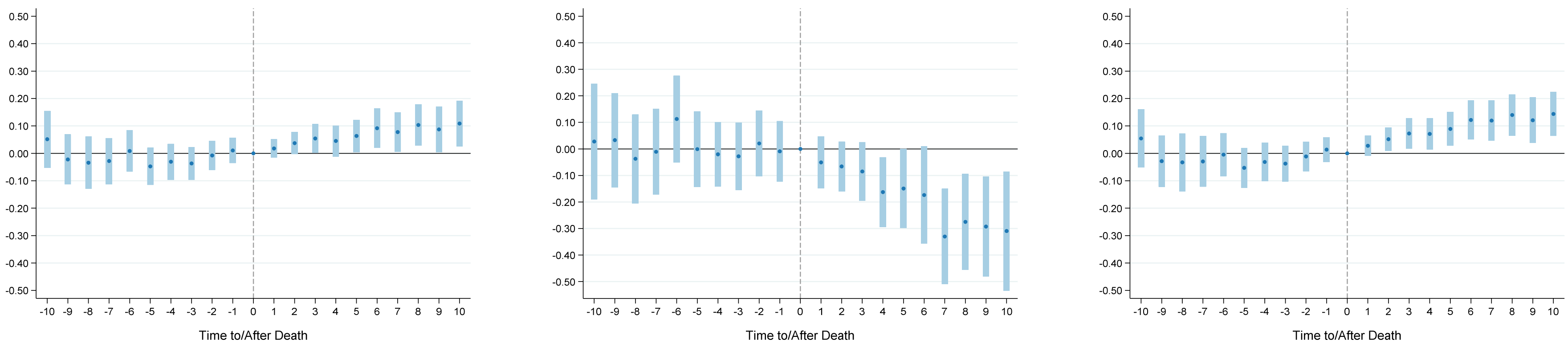
Figure F1
Timing of Death Events



Note: The number of distinct stars who die prematuraley during each year is indicated at the top of each bar.

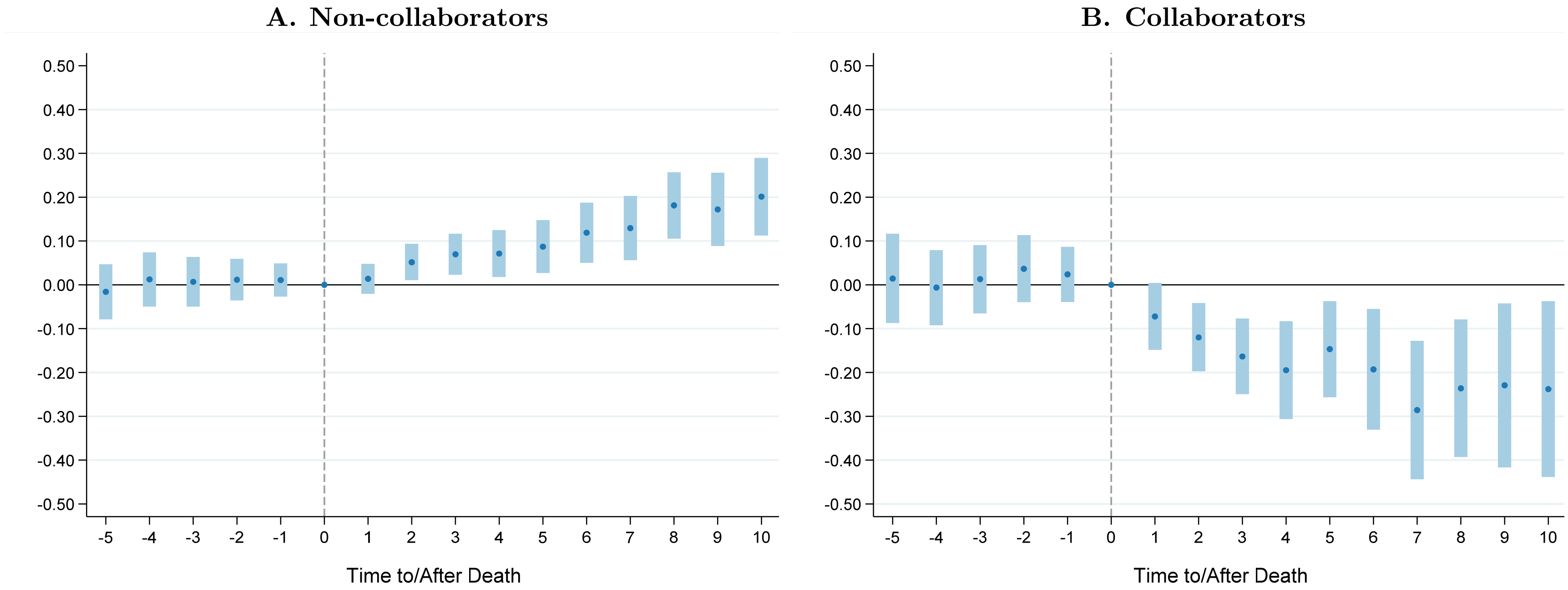
Figure F2
Effect of Star Scientist Death on Subfield Growth and Decline
Balanced Panel

A. **All Authors** B. **Collaborators** C. **Non-Collaborators**



<u>Note</u>: The graphs in this figure are patterned after Figure 2 in the main body of the manuscript. The dark blue dots correspond to coefficient estimates stemming from conditional (subfield) fixed effects Poisson specifications in which publication flows in subfields are regressed onto year effects, subfield age effects, as well as 20 interaction terms between treatment status and the number of years before/after the death event (the indicator variable for treatment status interacted with the year of death is omitted). The specifications also include a full set of lead and lag terms common to both the treated and control subfields to fully account for transitory trends in subfield activity around the time of the death. The sample used to estimate these specifications differs in one respect from our main sample: it has been extended from 2006 to 2012, which entails that at least nine years of data are available to identify the treatment effects far away from death (the latest date of death in our sample is 2003). The 95% confidence interval (corresponding to robust standard errors, clustered around star scientist) around these estimates is plotted with the vertical light blue lines.

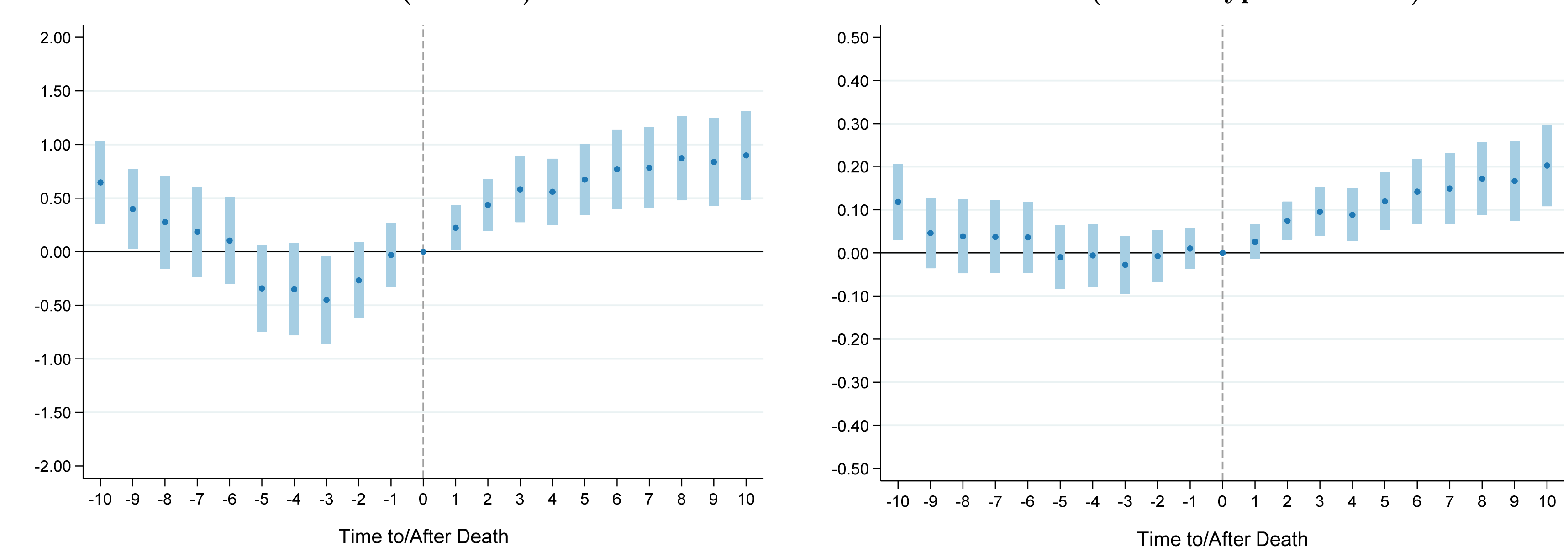
Figure F3: Effect of Star Scientist Death on Subfield Growth and Decline Aggregated up to the level of the star scientist



Note: The graphs in this figure are patterned after Panel B and C in Figure 2 in the main body of the manuscript. The dark blue dots correspond to coefficient estimates stemming conditional (star scientist) fixed effects Poisson specifications in which publication flows within the composite-subfield (comprising all the distinct related articles associated with a star's source articles) are regressed onto year effects, subfield age effects, as well as 20 interaction terms between treatment status and the number of years before/after the death event (the indicator variable for treatment status interacted with the year of death is omitted). The specifications also include a full set of lead and lag terms common to both the treated and control subfields to fully account for transitory trends in subfield activity around the time of the death. The 95% confidence interval (corresponding to robust standard errors, clustered around star scientist) around these estimates is plotted with the vertical light blue lines.

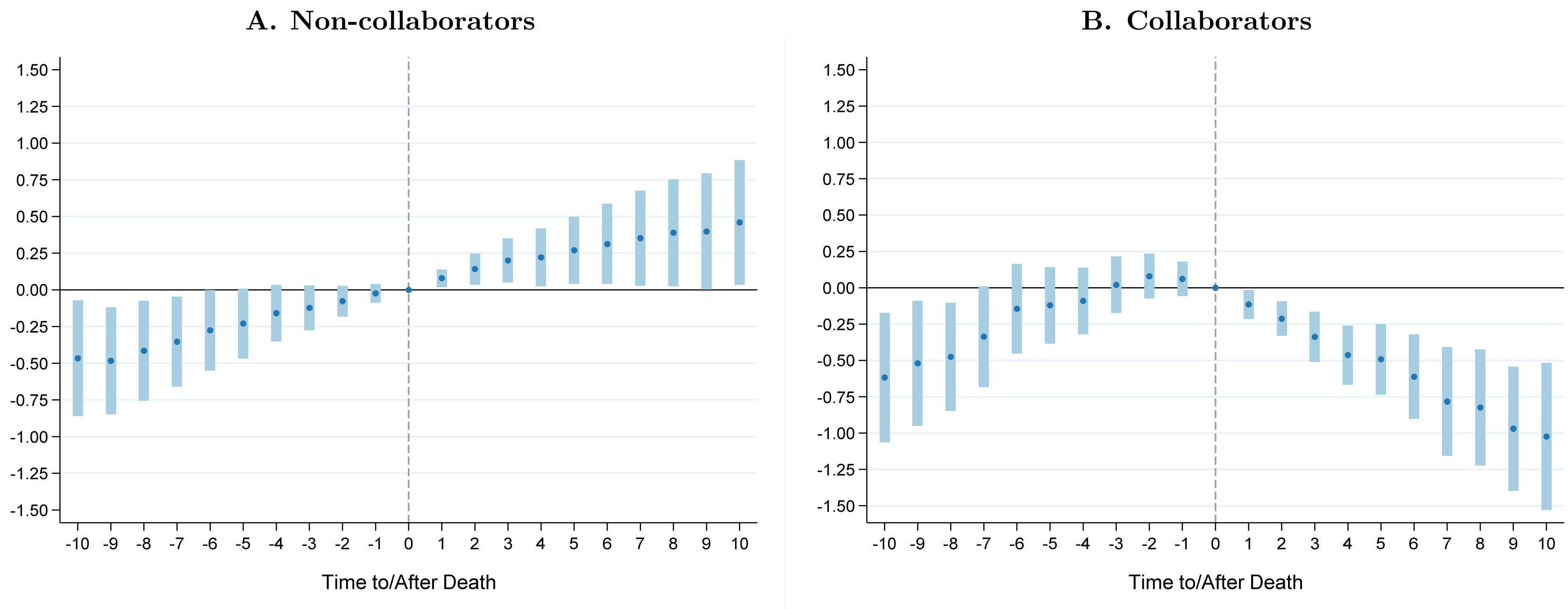
Figure F4: Effect of Star Scientist Death on Subfield Growth and Decline Non-Collaborators—Alternate Functional Forms

A. OLS (in levels) **B. OLS (inverse hyperbolic sine)**



Note: The graphs in this figure are patterned after Panel C in Figure 2 in the main body of the manuscript. The dark blue dots correspond to coefficient estimates stemming from subfield fixed effects OLS specifications in which publication flows by non-collaborators within a subfield are regressed onto year effects, subfield age effects, as well as 20 interaction terms between treatment status and the number of years before/after the death event (the indicator variable for treatment status interacted with the year of death is omitted). The specifications also include a full set of lead and lag terms common to both the treated and control subfields to fully account for transitory trends in subfield activity around the time of the death. In Panel A, the dependent variable is the "raw" count of articles in a subfield-year; In Panel B, these counts have been transformed using the inverse hyperbolic sine. The 95% confidence interval (corresponding to robust standard errors, clustered around star scientist) around these estimates is plotted with the vertical light blue lines.

Figure F5: Effect of Star Scientist Death on Subfield Growth and Decline
No Control Subfields



Note: The graphs in this figure are patterned after Panel B and C in Figure 2 in the main body of the manuscript. The dark blue dots correspond to coefficient estimates stemming from conditional (subfield) fixed effects Poisson specifications in which publication flows within a subfield are regressed onto year effects, subfield age effects, as well as 20 interaction terms between treatment status and the number of years before/after the death event (the indicator variable for treatment status interacted with the year of death is omitted). These regressions are run with subfield activity limited to non-collaborators of the star (Panel A), and with subfield activity limited to collaborators of the star (Panel B). The 95% confidence interval (corresponding to robust standard errors, clustered around star scientist) around these estimates is plotted with the vertical light blue lines

Appendix G: Displacement Effects

Conceptual challenges. We find that activity by non-collaborators of the star increases in the fields in which the superstar was active prior to his death. In principle, it is possible that commensurate declines can be observed in the fields where these related authors were active but the star was not. However, these displacement effects might be very diffuse—spread out over many subfields, and thus difficult to detect in our subfield-level of analysis. To examine this possibility more directly, we shift the level of analysis away from the subfield to that of the related author.

It is important to note however, that the panel dataset at the related author level is not simply the mirror image of the subfield panel dataset using an alternative way to aggregate the data. In particular, an author can only be represented in the sample if he was active in one of the star's subfields prior to his untimely death. But we have seen in Figure 3 and Table E10 that the bulk of the effect of death can be traced to new entrants in the subfield. We do not include these authors in the author-level analysis, because doing so would imply that the individuals are part of the sample because of an event that is itself a result of the treatment.

As a consequence, there should be no presumption that the magnitudes of the effect of star death at the author level and at the subfield level match. Since the author-level analysis necessarily excludes entrants, a reasonable conjecture is that the author-level effects will be smaller.

Author-level sample. In building up a sample of related authors, we face an important practical hurdle. A related author is frequently related to more than a single eminent scientist. Around which star should we anchor the analysis? In order to pin down a single year of treatment for each related author, we use two different metrics. The first is simply the number of related articles before the star's death—we associate to a related author the star with the highest count. The second metric is based on the cardinal relatedness score—we select the star that has the most highly related article among all the stars to whom the author is intellectually related. We proceed in a rigorously symmetric fashion for the related authors of control stars.

Since we are now choosing a focal star on which to anchor our analysis, but we know that authors are related to several distinct stars, we no longer maintain the distinction between those publications that are related and unrelated to a particular star. Rather, we turn our attention to the effect of superstar death on the total output of related authors (in terms of publications and NIH grants awarded). Recall that non-collaborators are contributing more within the subfields of the dead superstars with whom they are intellectually related (Table 3). Therefore, the absence of changes in total output would imply that this additional work is displacing work they were doing in other subfields, at least in part.

Results. We are now ready to proceed with a related author-level analysis whose structure parallels that of our main specifications at the subfield level. We investigate the effect of star death on related authors' (i) NIH grants awarded; (ii) publication output; and (iii) publication output split between "PI articles" and "non-PI articles."[xviii]

The results are displayed in Table G1. When looking at either publication or grant output, we do not find evidence of sustained increases after the death of a superstar. When focusing on authors associated with stars because of the number of related articles between the two, the effect of death tends to be small in magnitude and statistically indistinguishable from zero (the four leftmost columns of Table G1). These results change slightly when we focus on authors whose research was, at least in part, very closely related to

[xviii] PI articles—those where the focal author appears in first or last position on the authorship roster—are most intimately identified with his laboratory (Zuckerman 1968; Nagaoka and Owan 2014). In contrast, the articles where the related author appears in the middle of the authorship list correspond to research projects for which the author's substantive contribution might have been marginal.

that of the star. Here the magnitude of the effects are positive and relatively large in magnitude, but also imprecisely estimated.

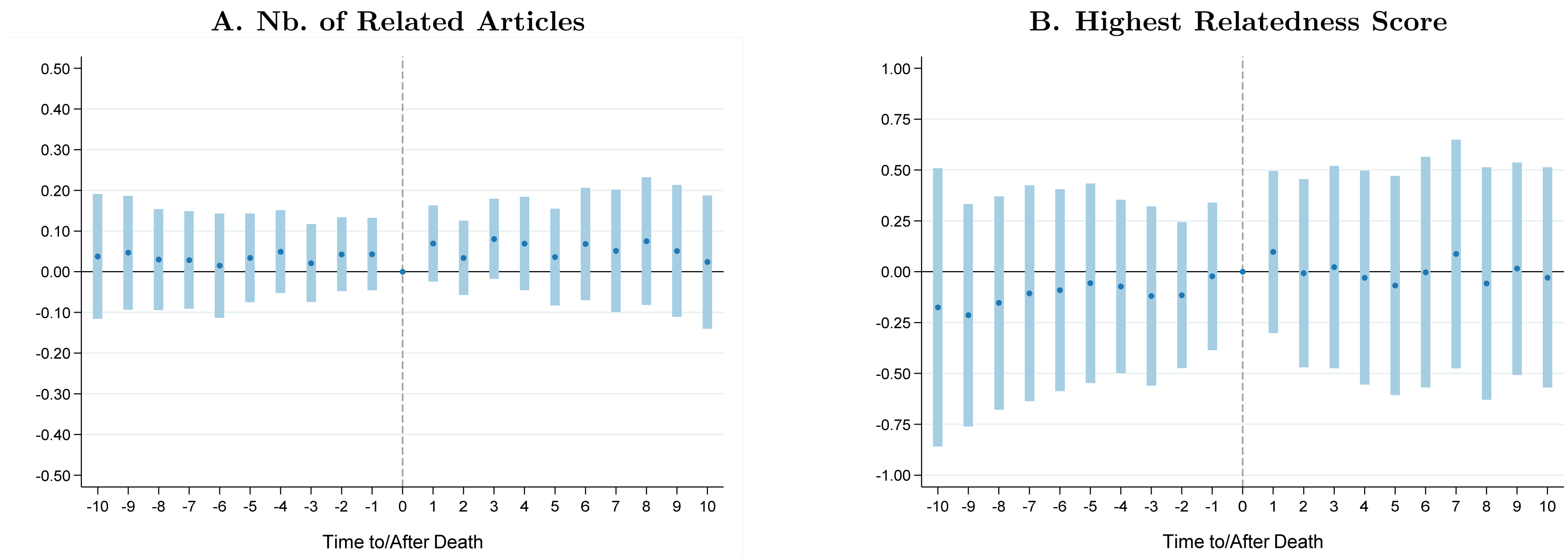
We estimate a dynamic version of these specifications and display the corresponding event study-style graphs in Figure G1 (publication output) and Figure G2 (grant output). In general, it appears from these figures that the total output of related authors neither expands nor contracts in the wake of a star's passing. Therefore, the related articles contributed to the star's subfields after they pass away most likely replace, at least in part, articles that these authors would have written in other intellectual domains had the star remained alive. Our results are therefore consistent with star extinction driving changes in the direction of scientific research, rather than shifting the overall level of scientific activity.

Table G1: Related Authors' Publication and Grant Output

	Nb. of Related Articles				**Highest Relatedness Score**			
	Nb. of NIH Grants	All Pubs.	PI Pubs.	Middle-Author Pubs.	Nb. of NIH Grants	All Pubs.	PI Pubs.	Middle-Author Pubs.
After Death	-0.020	0.003	0.014	-0.019	-0.019	0.092	0.087	0.072
	(0.022)	(0.060)	(0.083)	(0.055)	(0.053)	(0.228)	(0.319)	(0.172)
Nb. of Star Investigators	5,459	5,802	5,766	5,784	1,784	2,017	2,008	2,015
Nb. of Related Authors	26,728	44,649	42,654	43,483	2,944	3,850	3,811	3,840
Nb. of Star/Related Author Pairs	39,770	67,740	64,823	66,036	3,542	4,642	4,599	4,632
Nb. of Author-Year Obs.	888,746	1,402,293	1,357,179	1,382,976	94,132	120,918	120,249	120,822
Log Likelihood	-362,087	-772,285	-468,162	-595,167	-54,512	-86,098	-53,633	-71,209

Note: Estimates stem from conditional (related author) fixed effects Poisson specifications. The dependent variable is either the publication output for a related, non-collaborating author in a particular year, or the number of distinct NIH grants awarded to that author awarded in a particular year. In the four leftmost columns, each author is paired with the star with whom s/he had the highest number of related articles. In the four rightmost columns, each author is paired with the star with whom s/he had the related article with the highest relatedness score. All models incorporate a full suite of year effects and investigator age effects, as well as a term common to both treated and control authors that switches from zero to one after the death of the star. Exponentiating the coefficients and differencing from one yield numbers interpretable as elasticities. Robust standard errors in parentheses double-clustered at the level of the star & related authors. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$.

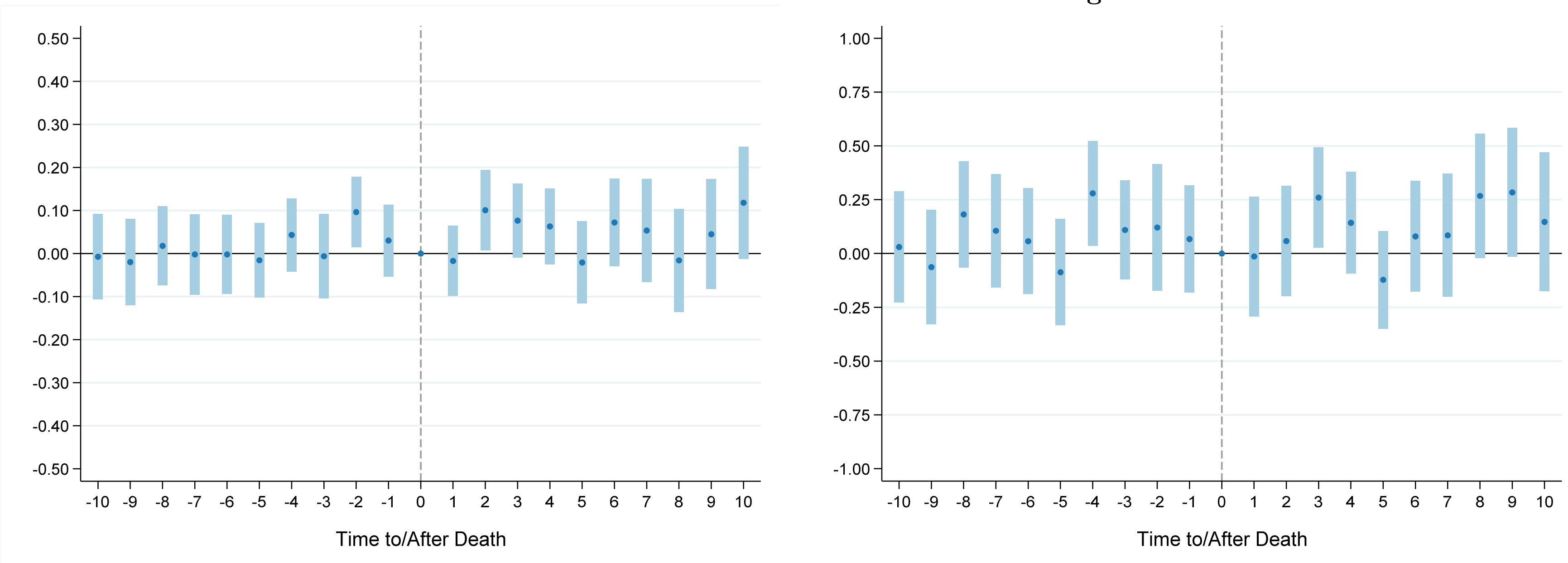
Figure G1: Effect of Star Scientist Death on Related Authors' Publication Output



Note: The dark blue dots in the above plots correspond to coefficient estimates stemming from conditional fixed effects specifications in which publication output for a related, non-collaborating author in a given year is regressed onto year effects, author age effects, as well as 20 interaction terms between treatment status and the number of years before/after the death event (the indicator variable for treatment status interacted with the year of death is omitted). The specifications also include a full set of lead and lag terms common to both treated and control authors. The 95% confidence intervals (corresponding to robust standard errors, clustered at the level of the associated star) around these estimates is plotted with the light-blue vertical lines; Panel A corresponds to a dynamic version of the specification in the second column of Table G1; Panel B corresponds to a dynamic version of the specification in the sixth column of Table G1.

Figure G2: Effect of Star Scientist Death on Related Authors' NIH Grants

A. Nb. of Related Articles **B. Highest Relatedness Score**



Note: The dark blue dots in the above plots correspond to coefficient estimates stemming from conditional fixed effects specifications in which the number of NIH grants awarded to a related, non-collaborating author in a given year is regressed onto year effects, author age effects, as well as 20 interaction terms between treatment status and the number of years before/after the death event (the indicator variable for treatment status interacted with the year of death is omitted). The specifications also include a full set of lead and lag terms common to both treated and control authors. The 95% confidence intervals (corresponding to robust standard errors, clustered at the level of the associated star) around these estimates is plotted with the light-blue vertical lines; Panel A corresponds to a dynamic version of the specification in the first column of Table G1; Panel B corresponds to a dynamic version of the specification in the fifth column of Table G1.

References

Aad, Georges et al. 2015. "Combined Measurement of the Higgs Boson Mass in *pp* Collisions at $\sqrt{s}$=7 and 8 TeV with the ATLAS and CMS Experiments." *Physical Review Letters* **114**(191803): 1-33.

Azoulay, Pierre, Andrew Stellman, and Joshua Graff Zivin. 2006. "PublicationHarvester: An Open-source Software Tool for Science Policy Research." *Research Policy* **35**(7): 970-974.

Azoulay, Pierre, Joshua Graff Zivin, and Gustavo Manso. 2011. "Incentives and Creativity: Evidence from the Academic Life Sciences." *RAND Journal of Economics* **42**(3): 527-554.

Bachrach, C. A., and Thelma Charen. 1978. "Selection of MEDLINE Contents, the Development of its Thesaurus, and the Indexing Process." *Medical Informatics (London)* **3**(3): 237-254.

Bhattacharya, Sanmitra, Viet Ha-Thuc, and Padmini Srinivasan. 2011. "MeSH: A Window Into Full Text for Document Summarization." *Bioinformatics* **27**(13): i120-i128.

Burbidge, John B., Lonnie Magee and A. Leslie Robb, 1988. "Alternative Transformations to Handle Extreme Values of the Dependent Variable." *Journal of the American Statistical Association* **83**(401): 123-127.

Blackwell, Matthew, Stefano Iacus, Gary King, and Giuseppe Porro. 2009. "`cem`: Coarsened Exact Matching in Stata." *The Stata Journal* **9**(4): 524-546.

Chu, S., J. DeRisi, M. Eisen, J. Mulholland, D. Botstein, P.O. Brown, and I. Herskowitz. 1998. "The Transcriptional Program of Sporulation in Budding Yeast." *Science* **282**(5389): 699-705.

Funk, Russell J., and Jason Owen-Smith. 2017. "A Dynamic Network Measure of Technological Change." *Management Science* **63**(3): 791-817.

Hausman, Jerry, Bronwyn H. Hall, and Zvi Griliches. 1984. "Econometric Models for Count Data with an Application to the Patents-R&D Relationship." *Econometrica* **52**(4): 909-938.

Jinek, Martin, Krzysztof Chylinski, Ines Fonfara, Michael Hauer, Jennifer A. Doudna, and Emmanuelle Charpentier. 2012. "A Programmable Dual-RNA-guided DNA Endonuclease in Adaptive Bacterial Immunity." *Science* **337**(6096): 816-821.

Law, John, and John Whittaker. 1992. "Mapping Acidification Research: A Test of the Co-word Method." *Scientometrics* **23**(3): 417-461.

Lin, Jimmy, and W. John Wilbur. 2007. "PubMed Related Articles: A Probabilistic Topic-based Model for Content Similarity." *BMC Bioinformatics* **8**(423): 1-14.

Myers, Kyle. 2018. "The Elasticity of Science." Working Paper*,* National Bureau of Economic Research.

Névéol, Aurélie, Rezarta Islamaj Dogan, and Zhiyong Lu. 2010. "Author Keywords in Biomedical Journal Articles." *AMIA Symposium Proceedings* 537-541.

Nagaoka, Sadao, and Hideo Owan. 2014. "Author Ordering in Scientifice Research: Evidence from Scientists Survey in the US and Japan." IIR Working Paper #13-23*,* Hitotsubashi University, Institute of Innovation Research.

Sopko, Richelle, Sheetal Raithatha, and David Stuart. 2002. "Phosphorylation and Maximal Activity of *Saccharomyces cerevisiae* Meiosis-Specific Transcription Factor Ndt80 Is Dependent on Ime2." *Molecular and Cellular Biology* **22**(20): 7024-7040.

Stephan, Paula E. 2012. *How Economics Shapes Science*. Cambridge, MA: Harvard University Press.

Whittaker, John. 1989. "Creativity and Conformity in Science: Titles, Keywords and Co-Word Analysis." *Social Studies of Science* **19**(3): 473-496.

Wilbur, W. John. 1998. "The Knowledge in Multiple Human Relevance Judgments." *ACM Transactions on Information Systems* **16**(2): 101-126.

Zuckerman, Harriet A. 1968. "Patterns of Name Ordering Among Authors of Scientific Papers: A Study of Social Symbolism and Its Ambiguity." *American Journal of Sociology* **74**(3): 276-291.