NBER WORKING PAPER SERIES

DOES SCIENCE ADVANCE ONE FUNERAL AT A TIME?

Pierre Azoulay
Christian Fons-Rosen
Joshua S. Graff Zivin

Working Paper 21788
http://www.nber.org/papers/w21788

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 2015

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). Fons-Rosen acknowledges financial support by the Spanish Ministry of Economy and Competitiveness (ECO2014-55555-P). 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, 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. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Does Science Advance One Funeral at a Time?
Pierre Azoulay, Christian Fons-Rosen, and Joshua S. Graff Zivin
NBER Working Paper No. 21788
December 2015, Revised October 2016
JEL No. I23,O31,O33

ABSTRACT

We study the extent to which eminent scientists shape the vitality of their fields by examining entry rates into the fields of 452 academic life scientists who pass away while at the peak of their scientific abilities. Key to our analyses is a novel way to delineate boundaries around scientific fields by appealing solely to intellectual linkages between scientists and their publications, rather than collaboration or co-citation patterns. Consistent with previous research, the flow of articles by collaborators into affected fields decreases precipitously after the death of a star scientist (relative to control fields). In contrast, we find that the flow of articles by non-collaborators increases by 8% 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. Overall, these results suggest that outsiders are reluctant to challenge leadership within a field when the star is alive and that a number of barriers may constrain entry even after she is gone. Intellectual, social, and re- source barriers all impede entry, with outsiders only entering subfields that offer a less hostile landscape for the support and acceptance of “foreign” ideas.

Pierre Azoulay
MIT Sloan School of Management
100 Main Street, E62-487
Cambridge, MA 02142
and NBER
pazoulay@mit.edu

Joshua S. Graff Zivin
University of California, San Diego
9500 Gilman Drive, MC 0519
La Jolla, CA 92093-0519
and NBER
jgraffzivin@ucsd.edu

Christian Fons-Rosen
Christian Fons-Rosen
Universitat Pompeu Fabra and CEPR
Barcelona GSE
Carrer Ramon Trias Fargas, 25-27
08005 Barcelona
SPAIN
christian.fons-rosen@upf.edu

"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

Knowledge accumulation—the process by which new research builds upon ideas developed in prior research—has been long understood to be of central importance to scientific progress and economic growth (Mokyr 2002). In deference to Sir Isaac Newton, this cumulative process is often referred to as "standing on the shoulders of giants," but is conceptualized more prosaically as the way in which researchers in one generation learn from and build upon prior research. Yet the literature is largely silent on the mechanisms that shape this slowly evolving process.[1]

What guides researchers when choosing between various approaches to study a given problem? Does science evolve according to autonomous laws, or is the direction of science influenced by individuals, incentives, and institutions? Philosophers and historians have long debated the extent to which the pragmatic success of a scientific theory determines how quickly it gains adherents, or its longevity (e.g., Kuhn [1970], Laudan [1977], and their many detractors). The epigraph of this paper encapsulates the jaundiced view, attributed to Planck, that the idiosyncratic stances of individual scientists can do much to alter, or at least delay, the course of scientific advance. Yet, the proposition that established scientists are slower than younger ones in accepting novel theories has received little empirical support whenever it has been put to the test (Hull et al. 1978; Gorham 1991; Levin et al. 1995). Moreover, in contrast to technology development where market forces shape the direction of research effort (however imperfectly, cf. Acemoglu [2012]), the choice of a problem-solving approach in basic research is less informed by market signals, and thus necessarily depends on a more nuanced system of non-pecuniary incentives (Feynman 1999; Foster et al. 2015).

[1] This stands in contrast to "paradigm shifts" (Kuhn 1970), which are exceedingly rare but garner far more scholarly attention. Bramoullé and Saint-Paul (2010) provide an equilibrium model of scientific revolutions with a Kuhnian flavor.

In this paper, we use a difference-in-differences setup to test "Planck's Principle" by examining how the premature death of 452 eminent academic life scientists alter the vitality (measured by publication rates and funding flows) of treated subfields in which these scientists actively published in the years immediately preceding their passing, compared to matched control subfields in which no eminent scientist dies. In contrast with prior work that focused on collaborators (Azoulay et al. 2010; Oettl 2012; Jaravel et al. 2015), our work leverages new tools to define scientific subfields in order to provide the first evidence on the response by non-collaborators. To our surprise, it is not competitors from within the field that assume the mantle, but rather outsiders 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 new directions that advance the frontier of knowledge within them. The rest of the manuscript tries to elucidate 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 and to those outside the club that once nucleated around the star. Though stars may have been a source of dynamism while alive, the turnover in leadership enables the injection of fresh ideas into the subfield, but only in those areas whose topology offers a less hostile landscape for the support and acceptance of "foreign" ideas.

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

[2]Considerable work outside of economics has examined the evolution of scientific fields through data visualization techniques (cf. Chavalarias and Cointet (2013) for a recent example). While interesting, this work has been largely descriptive and mostly silent regarding the behavioral mechanisms that might explain the birth, fusion, split, or death of scientific fields.

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 very detailed keyword information as well as the relative frequencies of these keywords in the scientific corpus.[3] As such we are able to define circumscribed areas of scientific inquiry that are independent of training, personal relations, or self-proclaimed areas of expertise.

In addition to providing evidence regarding a central question for scholars studying the scientific process, our paper is a departure for the field of the economics of science in that it can attend to the ways in which scientists position themselves simultaneously in an intellectual space as well as a social space, whose boundaries do not overlap (Borjas and Doran 2015). As such, our work can be understood as integrating the traditional concerns of economists—understanding how incentives and 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). These discoveries, in turn, have generated enormous health and welfare gains for economies around the world.

[3]Unlike in economics, keywords for all publications indexed by *PubMed* (most of the life sciences) are assigned by staff at the National Library of Medicine and are drawn from a controlled vocabulary thesaurus. Thus, concerns about strategic or endogenous keyword choices are minimized in this setting (cf. Appendix C for additional evidence on this point).

Second, the life science research workforce is exceedingly large and specialized. Academic medical centers in the United States employ 150,000 faculty members. 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 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 allows us to define scientific subfields in a way that is stripped of "social baggage" (the specifics of this process will be described in detail in Section 3.2).

Lastly, while accounts by practicing scientists indicate that collaboration plays a large role in both the creation and diffusion of new ideas (Reese 2004), 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 "ex-

tinction 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 the Institute of Medicine. Since these four 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. These include: (5) NIH MERIT awardees; (6) Howard Hughes Medical Investigators; and (7) early career prize winners. Appendix A provides additional details regarding these seven metrics of "superstardom."

For each scientist in the sample, 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]

[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).

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 will explore the sensitivity of our results to the age at death later). 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] Appendix G provides the full list of deceased superstars, together with their year of birth, year of death, institutional affiliation at the time of their passing, and a short description of their research expertise.

Table 1 provides descriptive statistics for the deceased superstar sample. The median star received his degree in 1957, died at 61 years old and was associated with 4 distinct subfields in the five years leading up to his/her death. 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 8,347 citations over the course of their careers (as of early 2014).

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 as inferred by the overlap in keywords between each article the star published in the five years prior to his/her death, and the rest of the scientific literature. Specifically, we use the *PubMed Related Citations Algorithm*

[5] We exclude from the sample one scientist who took his own life, and a further two for whom suicide could not be ruled out.

(PMRA) which relies heavily on Medical Subject Headings (MeSH). 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 scientific publication by professional indexers and not by the authors themselves.[6] We then use the "Related Articles" function in *PubMed* to harvest journal articles that are intellectually proximate to star scientists' own papers.[7]

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 C1, PMRA returns 72 original related journal articles for this source publication.[8] 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 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 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.

Consider now the second most-related article to Herskowitz's *Science* paper listed in Figure C1, "Phosphorylation and maximal activity of *Saccharomyces cerevisiae* meiosis-specific transcription factor Ndt80 is dependent on Ime2." Figure C2 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 "Phospho-

[6]The algorithm also uses as inputs title and abstract words, which are obviously selected by authors, rather than by NLM staff. However, neither the choice of MeSH keywords nor the algorithm depend on cited references contained in publications.

[7]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/`.

[8]Appendix C provides more details on the rules that govern the cut-off for the number of articles returned by PMRA for any given source article.

rylation..." is 94%, whereas the relatedness score for the most distant related article in Figure C1, "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.[9] 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 3,074 subfields, and 21,633 pairwise combination of subfields (we are only considering pairs of subfields associated with the same individual star). Appendix Figure C3 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 his/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. 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] 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 of the list.

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 s/he 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 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 (Borusyak and Jaravel 2016). 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.

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 an infrequent occurrence. 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 vs. non-collaborators of the star, or between those by intellectual outsiders vs. insiders, treated and control observations will always be defined with perfect symmetry.

3.4 Descriptive Statistics

The procedure described above yields a total of 34,216 distinct subfields; 3,074 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. Appendix Figure D2 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

[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.

can observe a great deal of overlap between the two histograms; the means are virtually identical, but the median is higher for control subfields (65) than for treated subfields (58). 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).

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 recipient 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

[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.

[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.

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 12% 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.

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 entry 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 is 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

[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.

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. Second, 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.

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 (real or placebo) 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. (2015) 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 (respectively placebo stars) inherit a counterfactual date of death because they are uniquely associated with a treated subfield (respectively deceased star) 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.

[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).

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 estimates based on the fixed-effect 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.

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 that appear in prestigious journals (Journal Impact Factor [JIF] higher than two) and those that appear in less prestigious journals (JIF lower than two); or 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 1 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 4.7%) 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 20.3% (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 publication activity by non-collaborators in the subfield increases by a statistically significant 7.9% (column 3).[18]

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 1 correspond to columns 1, 2, and 3 in Table 3).[19]

[18]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.

[19]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.

With treatment events staggered over time, a concern with these dynamic specifications 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. Could such heterogeneity confound 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 1 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.[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 1, 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. We explore further the distinction between sudden and anticipated events 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

[20]We quickly revert to studying the unbalanced sample for the rest of the manuscript, for three reasons. First, many of the covariates we need to explore heterogeneity of the treatment effect are not available after 2006. Second, though we can account precisely for the employment status of the control superstars up to 2006, some may retire, or even die in the years that follow. Third, the version of Figure 1 estimated in the unbalanced sample looks substantially the same as the one estimated in the balanced sample.

(counting 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.[21]

4.3 Understanding subfield growth patterns induced by a star's passing

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 entry into the field 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.

Article Characteristics. 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 focus on core issues, 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.[22]

Panel A of Table 4 produces a battery of estimates corresponding to each of these six bins in columns 2 through 7 (column 1 simply replicates the effect for all papers, regardless

[21] The event study graphs corresponding to the dynamics of funding flows are available from the authors, but also show close similarity to those displayed in Figure 1.

[22] 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 17,276,676 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 C2 (in Appendix C) received 39 citations from other articles in *PubMed* by 2014. This puts this article above the 76^{th} percentile of the citation distribution for articles published in 2002.

of impact, that was previously displayed in Table 3, column 3). A startling result is that the magnitude of the treatment effect increases sharply as 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.

Panels B and C 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. The results in column 1 imply that there is an important difference between the two type of events—subfield growth is more pronounced when the death of the star was anticipated. Decomposing this effect across the quantile bins as above reveals that these differences 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 contribution sharply—on the order of 40%. Because of this convergence in the upper tail, the remainder of the manuscript will lump together anticipated and unanticipated events.[23]

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 four different estimates. In both cases, the magnitude of the treatment effect pertaining to publication activity by proximate articles is approximately twice as large as the magnitude corresponding to more distant articles. These differences, however, are not themselves statistically significant at conventional levels. But we can at least rule out the conjecture that non-collaborators enter the field from the periphery. Their contributions seem to lie smack-dab in the middle of the subfield as it existed when the star was still alive.

[23]The most salient results reported below 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.

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. The first contains only publications that cite at least some work which belongs to the subfield identified by PMRA for the corresponding source. The second contains 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, as opposed to all other authors. 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. Finally, we investigate the vintage of the references cited by related articles. The last two columns in Panel B indicate that the new contributions are more likely to build on science of a more recent vintage.

Taken together, the results in Panels A and B of 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 questions—their MeSH keywords overlap with those of the source article even more than is typical for the "average" article in the subfield. 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.

Related Author Characteristics. 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. Table 6 reports these results. Perhaps the simplest author characteristic is age. For each related article in the subfield, we match the authorship roster to the AAMC Faculty Roster. Then, we compute the mean career age over matched authors for each related article. Since the median career age for matched authors turns out to be 16, we assign each article to one of two bins, the first comprising all related articles with an "older" authorship team (mean author career age greater than 16), the second comprising all related articles with a "younger" authorship team (mean author career age less than or equal to 16). We then compute publication activity separately for these two groups by aggregating these data up to the subfield/year level of analysis. As can be observed in the first two columns of Table 6, there really is not any difference in the magnitude of the post-death effect across these two groups.

The second step is to distinguish between the related articles with at least one eminent author from related articles for which none of the authors is particularly famous at the time of its publication. To do this, we use two distinct measures of eminence. The first is whether a matched author belongs to our sample of 12,935 stars. The second is whether a matched author belongs to an even more elite set comprising Nobel Prize winners, Howard Hughes Medical Investigators, and members of the National Academy of Sciences. In the final four columns to Table 6, we find that the effect is driven by related articles where none of the authors is particularly famous. This is consistent with the idea that non-elite scientists have stronger incentives to explore outside of their existing research trajectory, relative to more established scientists.

Finally, we examine the proximity in intellectual space between the 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 matched author, by computing the fraction of their publications that belongs to the star's subfields up to the year before the publication of each related article.[24] The distribution of this field overlap measure is displayed on Panel A of Figure 2, separately for authors on publications in the treated and control subfields. These distribution are extremely skewed, with a pronounced mass point at the origin: a full 50% of the related articles turn out to have authors with exactly zero intellectual overlap with the star's subfield. In addition to the bottom two quartiles, we create ten bins 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 top percentile bin. We then compute the corresponding measures of subfield activity by aggregating the data up to the subfield/year level. We opt to present the results graphically in Figure 2, Panel B. 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.[25]

[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.

[25]Table E4 in Appendix E corresponds to a simplified version of Figure 2 (with only four bins: below the median degree of field overlap, in the third quartile, in the top quartile but below the top ventile, and in the top ventile) presented as a table. Because some models failed to converge with the fixed-effects Poisson estimator, in both Figure 2 and Table E4, all coefficients stem from OLS regressions (with subfield fixed effects and the same set of controls as equation 1].

A striking pattern emerges. The authors driving the growth in 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. To borrow a term from industrial organization, they are new entrants into these subfields, though the evidence presented above also shows that they are not especially likely to be younger scientists overall.

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 of 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 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.[26]

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 (Li 2015; Brogaard et al. 2014), 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.[27] 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-

[26]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. An alternative is to report OLS specifications, but we prefer sticking with Poisson models estimated by Quasi-maximum likelihood because of the large number of zeros the outcome variable exhibits.

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

benefit calculations from entry by outside scholars toward delay or alternative activities. Table 7 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 stars 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 only to those subfields in which the most eminent among the stars were active, as measured by our citations measure in Table 7.[28]

Intellectual closure. 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. To explore the notion of "paradigmatic closure" as a barrier to field entry we develop two measures of intellectual coherence.

The 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

[28]More precisely, Table 8 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.

between zero and one) for the 25^{th} most related article in the subfield.[29] According to this metric, subfields exhibit wide variation in their degree of intellectual coherence, with a mean and median equal to 0.62 ($sd = 0.13$). The second index of intellectual coherence exploits the list of references cited in each article in the subfield before the star's death. We simply compute the proportion of these references that fall within the subfield. Our contention is that fields that are more self-referential will tend to dissuade outsiders from entering. Once again, we observe meaningful variation across subfields using this second index ($mean = 0.081; median = 0.067; sd = 0.059$).

Social closure. 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 this barrier we develop two additional measures of coherence, only in this case those designed to capture social cohesion rather than paradigmatic closure.

A natural way to capture endogamy within a subfield is to focus on the extent to which the star trained a large number of the junior scientists within it. We conjecture that the fields of stars who produced many intellectual "offspring" would be less welcoming to outsiders than those in which the stars did not train many graduate students or postdoctoral fellows. To identify trainees, we focus on the subset of coauthors 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 $\pm$ three years around the year in which the collaborator's highest degree was received. Our first index of social coherence at the subfield level is then simply the count of the number of investigators trained by the star before his/her (possibly counterfactual) death. Our second measure of social coherence summarizes the degree of subfield "cliquishness" 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.[30]

[29]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.

[30]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

Panel A of Table 8 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. In fact, in the subsamples of unusually coherent subfields, we find no statistical evidence of a publication influx after the passing of a star.[31]

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 lions 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.[32] 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 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 8 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

is the number of closed triplets over the total number of triplets (both open and closed, cf. Luce and Perry [1949]).

[31] We acknowledge that the *difference* between the estimates for more or less coherent subfields in unlikely to reach statistical significance at conventional levels. What seems notable, however, is that the magnitudes are consistently ordered across the measures we consider.

[32] 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 their appears to be a meaningful correlation between the number of editorials written and the propensity to be an editor.

rule, it appears more pronounced when stars have relatively few collaborators in influential positions. The results based on the fraction of funding in a field that is held by collaborators are even stronger. 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 Robustness checks and extensions

Appendix E presents results pertaining to robustness analyses and extensions. In Table E1, we probe the robustness of the core results presented in Table 3, Panel A after rolling up the data to the level of the star scientist (deceased or control). To do so, we simply proceed by lumping all related articles for each star together as if they belonged to a single subfield. The cost of this approach is that the sample of control stars that follows does not inherit an unambiguous date of death, since the same control star can act as control for subfields associated with different treated stars. As a result, the corresponding specifications do not include the $AFTER\_DEATH$ term common to controls and treated stars as in equation 1. Nevertheless, the results in Table E1 are very similar to those in Table 3, both in terms of magnitude and statistical significance. As explained in Section 3.2, we strongly prefer the subfield level of analysis, primarily because the subfields delineated by the *PubMed* Related Citations Algorithm exhibit very limited overlap (see Figure C3 in Appendix C).

The first three columns of Table E2 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 in the case of non-collaborators, but not

in the case of activity by collaborators: the treatment effect is positive and imprecisely estimated in this case. Moreover, the corresponding event study graphs (unreported) clarify 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 reason to add to the specification an additional level of difference—that provided by control subfields.

The last three columns of Table E2 display coefficients estimated by ordinary least squares, rather than the fixed effects Poisson model of Hausman et al. (1984). The pattern of coefficients is similar to that observed in Table 3 with respect to their sign and levels of statistical significance. The magnitudes cannot be readily compared, but can be reconciled easily. For instance, the mean of the dependent variable in the last column of Table E2 is 3.09; the point estimate implies that non-collaborators increase the level of their contribution to treated subfields, relative to control subfields by 0.335 article per year on average. This represents an increase of 10.8% at the mean of the data, which is close—but not identical—to the rate of increase estimated in the third column of Table 3 (7.9%).

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 (rather than simply reviews, editorials, or comments), 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 on Table E3 (which focuses only on publication activity in subfields by non-collaborators of the star), 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 choose to keep these older stars in the sample because a larger sample size affords us opportunities to explore mechanisms without losing power to detect nuanced effects statistically. The last two columns of Table E3 investigate whether a star's experience in the field (measured as the number of years between her first contribution in it and the year of death) moderates the core result. The median age in the field at the time of death is seven. We find no difference in the magnitude of the treatment effect along this dimension.

Effects on prior agenda. We find that non-collaborators of the star increase their publication activity in the fields in which the superstar was active prior to her death. Appendix F examines whether there is evidence of commensurate declines in publication activity for these related authors in the fields where they were active but the star was not. This investigation entails a change in the level of analysis, from the subfield level to the related author level. A practical difficulty is that a related author can be—and is in fact frequently—related to more than a single star. To get around this issue and pin down for each related author a single year of treatment and a clear demarcation between in-field and out-of-field output, we build a panel dataset of related authors and their publication output using two different methods. In the first method, we associate each related author with the star who died (possibly counterfactually) in the earliest year of all possible years of treatment. In the second method, we associate each related author with their most-related star (i.e., the star for whom the cardinal relatedness score between her source article and the author's related article is highest). Regardless of method, we divide each related author's output according to whether it belongs to one of the fields of the star with whom s/he is associated, or whether it belongs to none of these fields.[33] Table F1 then examines how these measures of output shift after the death event, relative to before, for treated authors, relative to control authors. We also distinguish between the overall number of publications, the number of publications where the related author is in the middle of the authorship roster, and the number of publications where the related author is pivotal (in either first or last authorship position).

Panel A corresponds to the results obtained following the "earliest treating star" method. Panel B corresponds to the results obtained following the "most-related treating star" method. Regardless of the method employed, some stable patterns emerge. We can detect large effects on the rate of production of in-field articles, consistent with the results obtained when performing our analysis at the subfield level. Conversely, the magnitudes for the treatment effect on out-of-field output are typically much smaller, imprecisely estimated, and not always of the same sign. Figure F1 presents the corresponding event-study graphs (only for out-of-field publication output). The main takeaway is that we cannot detect any evidence of displacement. Non-collaborating related authors appear to increase their overall output modestly in the wake of a superstar's premature passing.[34]

[33]Since the universe of fields in large, it is important to note that our simple counts of activity outside of a star's fields will not be able to detect moderate shifts in foci *within* that set of activities.

[34]The absence of displacement is reassuring from one standpoint: if authors and indexers were just changing keywords opportunistically to position themselves strategically with respect to subfields that have lost their stars, but without really changing the content of their investigations, then we would expect to observe

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 scientist exerts on the dynamics of growth—or decline—for the fields in which s/he 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

decreases in out-of-field activity commensurate with the increase in within-field contributions. This does not appear to be the case.

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. Example 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).

Our work leaves many questions unanswered. What is the fate of the fields that these new entrants departed? Do they decay, or instead "merge" with those whose star departed prematurely? Given a finite supply of scientists and the adjustment costs involved in switching scientific focus, one would expect some other field to contract on the margin in the wake of a superstar's passing. Is this marginal field more novel, or already established? We are pursuing these questions in ongoing work.

References

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.

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

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

Borusyak, Kirill, and Xavier Jaravel. 2016. "Revisiting Event Study Designs." Working Paper, Harvard University. Available at SSRN: `http://ssrn.com/abstract=2826228`.

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.

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.

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.

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

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. 2015. "Team-Specific Capital and Innovation." Working Paper, Harvard University.

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.

Laudan, Larry. 1998. *Progress and its Problems: Towards a Theory of Scientific Growth*. Berkeley, CA: University of California 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. 2016. "Expertise vs. Bias in Evaluation: Evidence from the NIH." Forthcoming, *American Economic Journal: Microeconomics*.

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.

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.

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

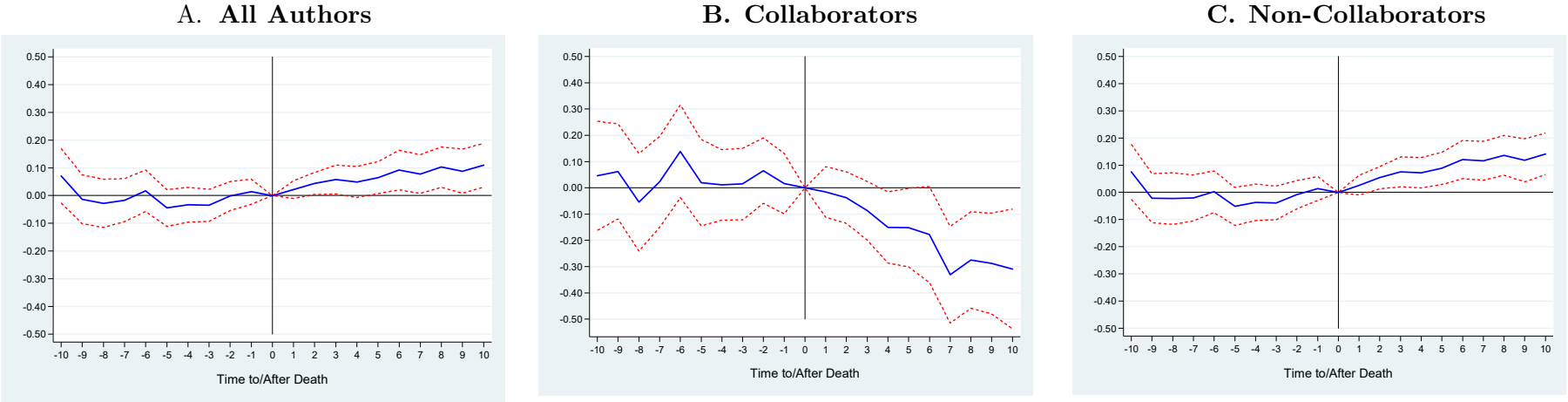
Reese, Thomas S. 2004. "My Collaboration with John Heuser." *European Journal of Cell Biology* **83**(6): 243-244.

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

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

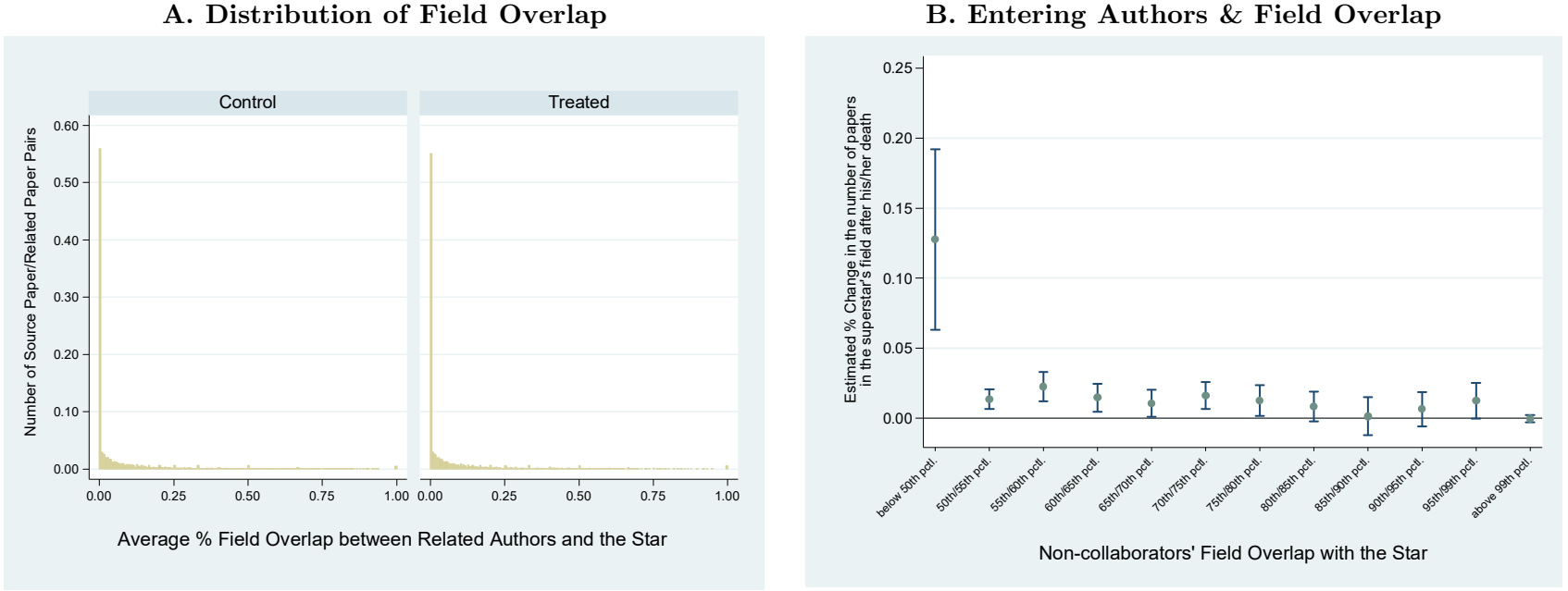
Wooldridge, Jeffrey M. 1997. "Quasi-Likelihood Methods for Count Data." In M. Hashem Pesaran, and Peter Schmidt (Eds.), *Handbook of Applied Econometrics*, pp. 352-406. Oxford: Blackwell.

Figure 1
Effect of Star Scientist Death on Subfield Growth and Decline



Note: The solid blue lines 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 sample used to estimate these specifications differs in one respect from our main sample: it has been extended from 2006 to 2012, which ensures 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). Our main sample stops the observation window in 2006, since many of the covariates needed to generate the estimates in Tables 5 through 8 are not available after 2006. When the analysis is restricted to the years 1970-2006 (i.e., with an unbalanced sample), the event study graphs look substantially similar to those above. The 95% confidence interval (corresponding to (QML) robust standard errors, clustered around star scientist) around these estimates is plotted with dashed red 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 2
Characteristics of Related Authors: Competitors or Outsiders?



Note: Panel A displays the distribution of field overlap for every combination of source article (authored by the star in the 5 years before her passing) and related article. We create a metric of intellectual proximity for each matched author on a related article, by computing the fraction of their publications that belongs to the star's subfield up to its year of publication. Whenever we can match more than one related author to the AAMC Faculty Roster on a given article, it is the most proximate scientist on the authorship roster which determines the particular bin within which an article falls. 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. 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. Appendix Table E4 displays these results in regression form.

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	6.801	4	7.298	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	69.638	65	36.780	1	216
Baseline Stock of Related Articles in the Field, Non-Collaborators	61.406	57	33.459	1	208
Baseline Stock of Related Articles in the Field, Collaborators	8.232	7	6.946	0	77
Source Article Nb. of Authors	3.970	3	1.793	1	15
Source Article Citations at Baseline	16.333	6	28.066	0	354
Source Article Long-run Citations	70.437	46	93.274	1	1505
Investigator Gender	0.067	0	0.167	0	1
Investigator Year of Degree	1960.546	1962	10.920	1926	1989
Death Year	1991.125	1991	7.970	1975	2003
Age at Death	58.100	58	8.796	34	91
Years of Experience in the Field	8.162	8	4.290	0	37
Investigator Cumulative Nb. of Publications	164	142	100	1	861
Investigator Cumulative NIH Funding at Baseline	$18,783,603	$14,291,121	$20,016,978	0	$220,856,880
Investigator Cumulative Nb. of Citations	12,143	9,897	9,996	9	143,383
Treated Subfields (N=3,074)					
Baseline Stock of Related Articles in the Field	69.398	58	46.644	0	225
Baseline Stock of Related Articles in the Field, Non-Collaborators	61.250	50	43.282	0	219
Baseline Stock of Related Articles in the Field, Collaborators	8.148	5	8.850	0	62
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
Years of Experience in the Field	8.392	7	5.915	0	38
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: Main Effect of Superstar Death

	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	0.046†	-0.228**	0.076**	0.043	-0.253**	0.105**
	(0.026)	(0.057)	(0.026)	(0.034)	(0.076)	(0.032)
Nb. of Investigators	6,260	6,111	6,260	6,213	5,673	6,200
Nb. of Fields	34,211	33,094	34,211	33,891	29,175	33,785
Nb. of Field-Year Obs.	1,258,911	1,217,857	1,258,911	1,049,285	903,224	1,046,021
Log Likelihood	-2,722,190	-712,808	-2,602,697	-1,312,147	-464,353	-1,186,098

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 $100\times(\exp[0.076]-1)=7.90\%$. 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. This is also true for the results reported in Tables 4 through 8.

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

Table 4: Breakdown by Long-run Citation Impact [Non-collaborators Only]

	All Pubs	Bttm. Quartile	2nd Quartile	3rd Quartile	Btw. 75th and 95th pctl.	Btw. 95th and 99th pctl.	Above 99th pctl.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: All causes of death							
After Death	0.076**	-0.051	-0.005	0.020	0.112**	0.226**	0.322**
	(0.026)	(0.034)	(0.031)	(0.029)	(0.033)	(0.047)	(0.076)
Nb. of Investigators	6,260	6,202	6,259	6,258	6,256	6,149	5,256
Nb. of Fields	34,211	33,353	34,197	34,206	34,201	33,161	21,532
Nb. of Field-Year Obs.	1,258,911	1,227,580	1,258,399	1,258,726	1,258,541	1,220,256	792,238
Log Likelihood	-2,602,697	-547,597	-1,063,216	-1,386,147	-1,427,762	-521,702	-148,477
Panel B: Anticipated							
After Death	0.100**	0.004	0.044	0.060	0.123**	0.173**	0.305**
	(0.034)	(0.046)	(0.039)	(0.037)	(0.044)	(0.066)	(0.109)
Nb. of Investigators	4,018	3,964	4,017	4,016	4,014	3,936	3,199
Nb. of Fields	15,084	14,744	15,079	15,082	15,076	14,601	9,434
Nb. of Field-Year Obs.	554,867	542,462	554,682	554,793	554,571	537,074	346,942
Log Likelihood	-1,162,200	-250,697	-478,506	-615,584	-626,319	-225,876	-64,011
Panel C: Sudden							
After Death	0.046	-0.106*	-0.049	-0.028	0.094†	0.275**	0.361**
	(0.042)	(0.051)	(0.049)	(0.048)	(0.052)	(0.070)	(0.110)
Nb. of Investigators	4,656	4,594	4,656	4,656	4,656	4,588	3,754
Nb. of Fields	17,543	17,041	17,534	17,540	17,542	17,045	11,188
Nb. of Field-Year Obs.	645,545	627,203	645,218	645,434	645,508	627,233	411,686
Log Likelihood	-1,307,660	-268,188	-528,480	-699,583	-733,600	-274,086	-79,021

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, 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*. Panel B and Panel C present the same specifications, but run on two distinct subsamples: In Panel B, the 1,576 subfields associated with 229 stars whose death is anticipated (along with the corresponding control subfields); and in Panel C, the 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 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 $100\times(\exp[0.076]-1)=7.90\%$.

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

Table 5: Breakdown by Intellectual Proximity to the Work of the Star [Non-collab. Only]

Panel A	**All Pubs**	**Cardinal Measure**		**Ordinal Measure**	
		Intllct. Proximate Articles	Intllct. Distant Articles	Intllct. Proximate Articles	Intllct. Distant Articles
After Death	0.076**	0.109**	0.059*	0.122**	0.061*
	(0.026)	(0.032)	(0.027)	(0.029)	(0.027)
Nb. of Investigators	6,260	6,101	6,214	6,258	6,260
Nb. of Fields	34,211	30,576	33,780	34,187	34,211
Nb. of Field-Year Obs.	1,258,911	1,124,963	1,243,169	1,258,023	1,258,911
Log Likelihood	-2,602,697	-880,290	-2,259,464	-1,082,892	-2,303,024

Panel B	**In-field vs. Out-of-field References**		**Backward Citations to the Star's Bibliome**		**Average Backward Citation Lag**	
	w/ in-field references	w/o in-field references	w/ references to the star	w/o references to the star	Below Median	Above Median
After Death	0.026	0.103**	0.066†	0.146**	0.070*	-0.010
	(0.030)	(0.028)	(0.034)	(0.030)	(0.034)	(0.029)
Nb. of Investigators	6,260	6,257	6,245	6,259	6,260	6,259
Nb. of Fields	34,209	34,194	34,167	34,137	34,208	34,208
Nb. of Field-Year Obs.	1,258,837	1,258,290	1,257,297	1,256,200	1,258,810	1,258,816
Log Likelihood	-1,817,724	-1,714,995	-1,825,511	-1,659,982	-1,805,312	-1,691,725

Note: Estimates stem from conditional (subfield) fixed effects Poisson specifications. In Panel A, the dependent variable is the total number of publications 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 (the first column of the table merely replicates the estimate already found in Table 3, column (3), for comparison purposes). For the cardinal measure, a related article is deemed proximate if its similarity score is above .70, which corresponds to the top quartile of similarity in the sample. For the ordinal measure, a related article is deemed proximate if its similarity rank is below 40, which also corresponds to the top quartile of similarity in the sample. In Panel B, we separate the related articles by examining the type of references cited in their bibliography. Each cited reference can be either in the source's PMRA field, or outside of it; it can be a publication of the star scientist, or of someone else's; and the average lag between the related article's publication year and that of the articles it cites can be either above or below the median (6.5 years). 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 $100\times(\exp[0.076]-1)=7.90\%$.

Robust (QML) 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 Related Author Characteristics [Non-collaborators Only]

	Author Career Age		Star Author		Elite Author	
	> 16	≤ 16	With	Without	With	Without
After Death	0.091**	0.095**	0.049†	0.094**	0.032	0.076**
	(0.028)	(0.031)	(0.027)	(0.034)	(0.055)	(0.026)
Nb. of Investigators	6,260	6,259	6,260	6,253	5,582	6,260
Nb. of Fields	34,207	34,207	34,211	34,151	27,197	34,211
Nb. of Field-Year Obs.	1,258,763	1,258,763	1,258,911	1,256,714	1,001,131	1,258,911
Log Likelihood	-1,282,393	-1,418,924	-2,190,045	-1,229,336	-279,203	-2,556,867

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, where these publications have scientists on their authorship roster with certain demographic characteristics. The first two columns examine the impact of related author age. Hence, we compute the average career age of every author we could match with the AAMC Roster, and compute the average age of the authorship team for the related article, at the time of its publication. We then divide related articles according to whether the average career age for identified authors is above or below 16 (the median in our sample), and we aggregate up our measure of subfield activity separately for these two groups. We proceed similarly for the middle two columns (whether or not a related article has one of our 12,935 stars on its authorship roster) and for the last two columns (whether or not a related article has a member of the NAS or an HHMI investigator on its authorship roster). 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-collaborators after the superstar passes away—a statistically significant $100\times(\exp[0.091]-1)=9.53\%$.

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

Table 7: Breakdown by Star Scientist Characteristics [Non-collaborators Only]

	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	0.054	0.103*	0.030	0.119**	0.015	0.137**	0.058*	0.128**
	(0.033)	(0.045)	(0.037)	(0.036)	(0.035)	(0.048)	(0.027)	(0.040)
Nb. of Investigators	2,901	4,836	2,792	4,619	3,047	4,288	5,022	4,462
Nb. of Fields	17,208	17,003	17,327	16,884	15,726	15,485	16,967	17,244
Nb. of Field-Year Obs.	632,010	626,901	636,708	622,203	578,087	570,590	624,449	634,462
Log Likelihood	-1,284,625	-1,315,748	-1,273,531	-1,324,678	-1,184,870	-1,185,724	-1,340,763	-1,222,805

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. 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 of implies that treated subfields see an increase in the number of contributions by non-collaborators after the superstar passes away—a statistically significant $100\times(\exp[0.103]\text{-}1)=10.85\%$.

Robust (QML) 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 Nature of Entry Barriers

Panel A

	"Intellectual" Subfield Coherence				"Social" Subfield Coherence			
	PMRA-based definition		Citation-based definition		Nb. of Trainees		Cliquishness	
	Below Median	Above Median	Below Median	Above Median	Below Median	Above Median	Below Median	Above Median
After Death	0.176**	0.063	0.102†	0.086*	0.114*	0.023	0.120*	0.063
	(0.036)	(0.042)	(0.054)	(0.039)	(0.045)	(0.057)	(0.049)	(0.046)
Nb. of Investigators	3,355	3,202	3,420	3,113	3,248	1,903	2,859	3,557
Nb. of Fields	9,068	7,816	8,638	8,246	8,607	8,277	8,034	8,850
Nb. of Field-Year Obs.	334,357	287,846	318,247	303,956	316,937	305,266	296,320	325,883
Log Likelihood	-689,898	-636,190	-654,563	-677,479	-677,872	-667,756	-652,392	-681,321

Panel B

	Indirect Control through Collaborators					
	Editorial Channel		NIH Study Section Channel		Fraction of Field NIH Funding	
	Below Median	Above Median	Below Median	Above Median	Below Median	Above Median
After Death	0.139**	0.079†	0.139**	0.101*	0.163**	0.069
	(0.051)	(0.042)	(0.049)	(0.041)	(0.047)	(0.049)
Nb. of Investigators	3,452	2,068	3,396	2,264	3,559	2,521
Nb. of Fields	11,105	5,779	10,441	6,443	9,861	7,023
Nb. of Field-Year Obs.	409,839	212,364	385,123	237,080	363,601	258,602
Log Likelihood	-911,698	-426,446	-845,335	-474,501	-807,292	-528,220

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 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 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. 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 $100\times(\exp[0.139]-1)=14.91\%$.

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

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 either 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.

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, 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]

[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.

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

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.

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.

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

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])`.

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

Traditionally, it has been very difficult to assign to individual scientists, or articles, a fixed address in "idea space," and this data constraint explains in large part why bibliometric analyses typically focus on the determinants of the rate of scientific progress rather than its direction. The empirical exercise in this paper hinges crucially on the ability to relax this constraint in a way that is consistent across death events and also requires little, if any, human judgement.

This challenge is met here by the use of the *PubMed* Related Citations Algorithm [PMRA], a probabilistic, topic-based model for content similarity that underlies the "related articles" search feature in *PubMed*. This database feature is designed to help a typical user search through the literature by presenting a set of records topically related to any article returned by a *PubMed* search query.[vi] To assess the degree of intellectual similarity between any two *PubMed* records, PMRA relies crucially on MeSH keywords. MeSH is the National Library of Medicine's [NLM] controlled vocabulary thesaurus. It consists of sets of terms arranged in a hierarchical structure that permit searching at various levels of specificity. There are 27,149 descriptors in the 2013 MeSH edition. 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). What is key for our purposes is that the subjectivity inherent in any indexing task is confined to the MeSH term assignment process and does not involve the articles' authors.[vii]

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*

[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])`

[vi] Lin and Wilbur (2007) report that one fifth of "non-trivial" browser sessions in *PubMed* involve at least one invocation of PMRA.

[vii] This is a slight exaggeration: PMRA also makes use of title and abstract words to determine the proximity of any two pairs of articles in the intellectual space. These inputs are obviously selected by authors, rather than by NLM staff. However, neither the choice of MeSH keywords nor the algorithm depend on cited references contained in publications.

lengths. To carry out such a program, one must first define what a word is. For us, a word is basically an 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.[viii]

The algorithm uses a cut-off rule to determine the number of related citations 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 A is related to Publication B, Publication B must also be related to publication A. As a result, the set of related citations for a given source article may contain many more than 100 publications.[ix]

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).

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

[ix] The effective number of related articles returned by PMRA varies between 58 and 2,097 in the sample of 3,074 source articles published by the 452 star scientists in the five years preceding their death. The mean is 185 related articles, and the median 141.

An Example. In Figure C1, 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.[x] 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 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 C2 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 C1, "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 implication of this 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 3,074 subfields, and 21,633 pairwise combination of subfields (we are only considering pairs of subfields associated with the same individual star). Figure C3 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.

Endogenous keyword choice. PMRA is a modern implementation of *co-word analysis*, a content analysis technique that uses patterns of co-occurrence of pairs of items (i.e., title words or phrases, or keywords) in a corpus of texts to identify the relationships between ideas within the subject areas presented in these texts (Callon et al. 1991; He 1999). One long-standing concern among practitioners of this technique has been the

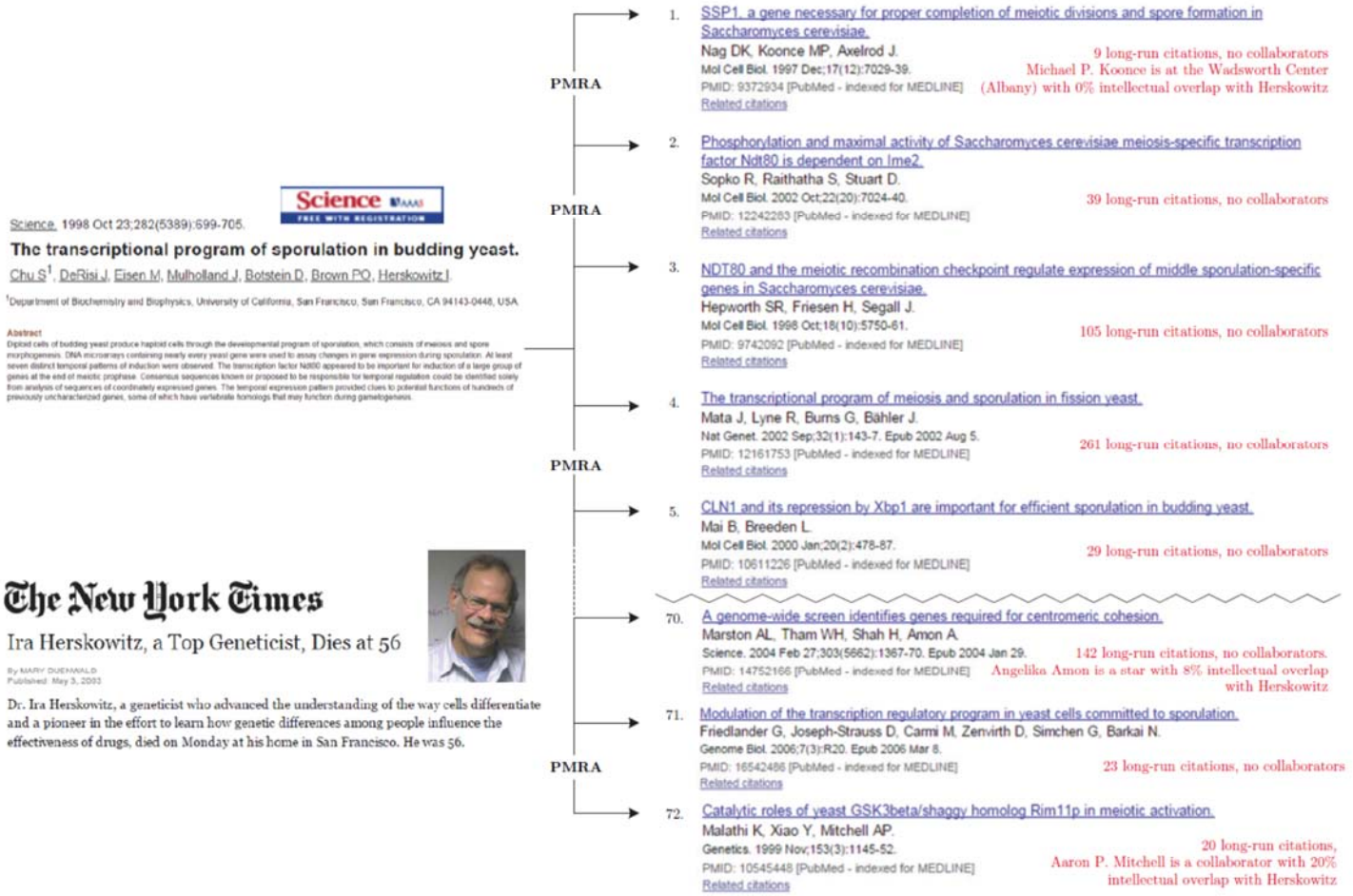
[x] Why exactly 72? In fact, PMRA lists 152 "intellectual neighbors" for PubMed ID 9784122. But once we exclude articles published after 2006 (the end of our observation period), purge from the list reviews, editorials and other miscellaneous "non-original" content, and drop a handful of articles that 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.

"indexer effect" (Whittaker 1989). Clustering algorithm 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?

Indexer effects could manifest themselves in three distinct ways. First, indexers may have available a lexicon of permitted keywords which is itself out of date. Second, there is an inevitable delay between the publication of an article and the appearance of an entry in *PubMed*. Third, indexers, in their efforts to be helpful to users of the database, may use combinations of keywords which reflect the conventional views of the field. The first two concerns are legitimate, but probably have only a limited impact on the accuracy of the relationships between articles which PMRA deems related. This is because the NLM continually revises and updates the MeSH vocabulary, precisely in an attempt to neutralize keyword vintage effects. Moreover, the time elapsed between an article's publication and the indexing task has shrunk dramatically, though time lag issues might have been a first-order challenge when MeSH was created, back in 1963. The last concern strikes us as being potentially more serious; a few 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).

In general, indexer effects are not a concern for our identification strategy because one would expect them to affect treated and control subfields equally. But what if the death of a superstar scientist, in and of itself, changes the mix of keywords that is used to describe a given set of scientific phenomena, research questions, or experimental procedures? To some extent, this is an empirical question. We merge every article in the subfields in our data to the set of MeSH keywords that were assigned to them, and we calculate, for each keyword, its vintage, i.e., the year in which it was first introduced in the scientific literature (i.e., within the *PubMed* universe). With this information, we then compute the average age of MeSH keywords in a subfield in a given year, and we can reprise the main analysis in the paper with this new outcome variable. Figure C4 below displays the event study graph (analogous to the one presented in Figure 1, Panel A) using average MeSH age as the outcome of interest. The point estimates are minuscule, and there is no evidence of keyword renewal after the death, relative to before the death. The same patterns hold when focusing on the median age of keywords, or the age of the youngest keywords in a subfield. We also replicate Figure C4 while limiting the articles to the set of articles by collaborators and non-collaborators separately. In all cases, we obtain a flat profile for the path of the outcome. This provides suggestive evidence that professional indexers are not influenced by passing of famous authors, at least as a first-order approximation.

Figure C1: From Source to Related Articles

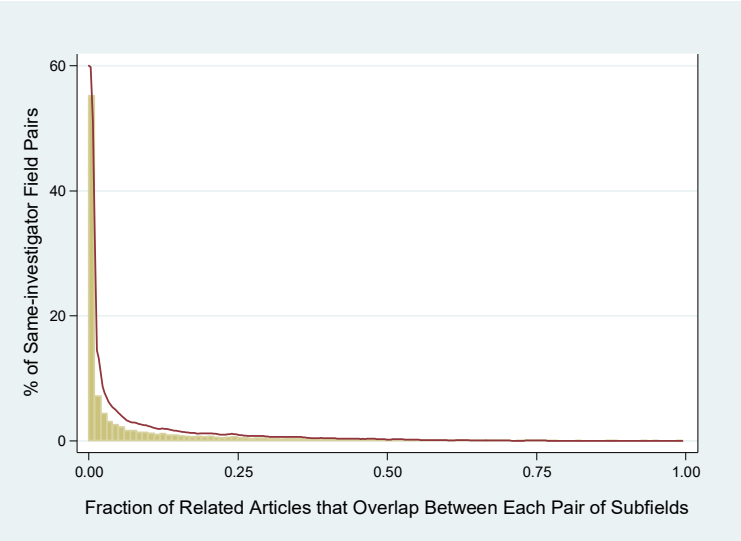


Note: We illustrate the process of identifying the related articles through the use of an example. Ira Herskowitz, a superstar scientist in our sample, died in 2003. In the five years prior to his death (1998-2002), Herskowitz was the last author on 12 publications. One of these publications is *"The 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.

Figure C2: PMRA and MeSH Term Overlap–An Example

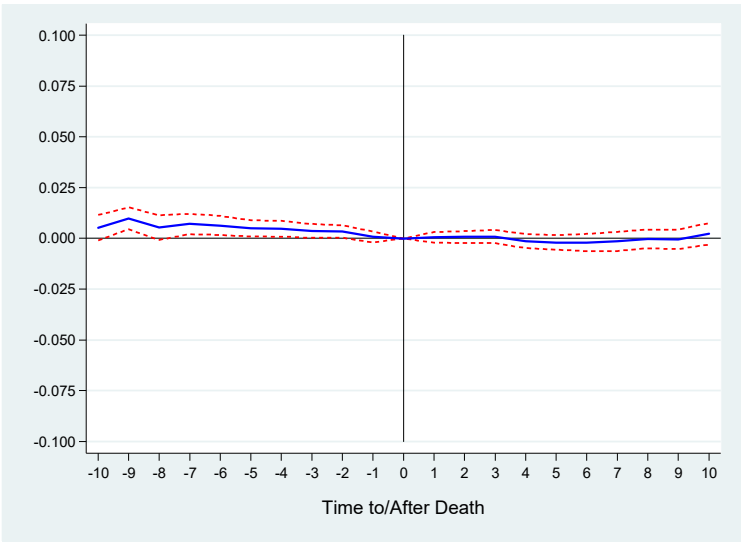
Source Article	PMRA-Linked Article
Chu et al., "The transcriptional program of sporulation in budding yeast." *Science*, 1998.	Sopko et al. "Phosphorylation and maximal activity of Saccharomyces cerevisiae meiosis-specific transcription factor Ndt80 is dependent on Ime2." *MCB*, 2002.
PMID #9784122	**PMID #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

Figure C3: Within-star Pairwise Subfield Overlap



Note: We compute the fraction of articles that overlap between every pair of subfields in which a deceased star is active in the five years leading to his/her death. There are 21,633 subfield pairs corresponding to the 3,074 distinct subfields for the 452 deceased superstars. The median degree of overlap between subfield pairs is 0, and the mean is 0.06. Subfields that overlap by 50% or more belong to the top two percentiles of the pairwise overlap distribution.

Figure C4: Within-star Pairwise Subfield Overlap



Note: The solid blue line corresponds to coefficient estimates stemming from conditional (subfield) fixed effects Poisson specifications in which average MeSH keyword age 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 (QML) robust standard errors, clustered around star scientist) around these estimates is plotted with dashed red lines.

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

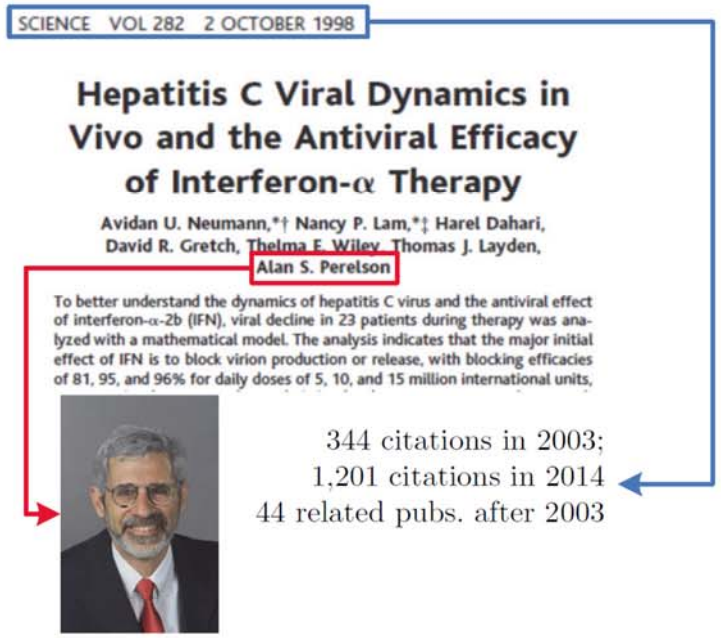
of the star author on the authorship roster (only last authorship position is considered); journal; and year of publication. The first three covariates only need to match within relatively coarse bins. For instance, we create nine career age categories: less than 10 years; between 10 and 20 years; between 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.

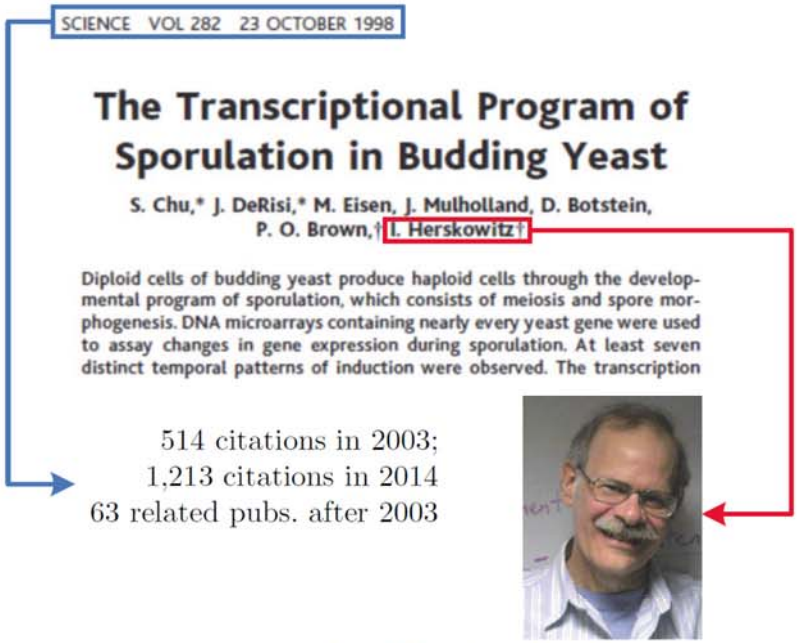
We match approximately 75% of the treated source articles in this way. Some further trimming of the control articles is needed. First, we eliminate any control that shares any author with the treated source. Second, we eliminate any control article with a dead star scientist on its authorship roster, even if s/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



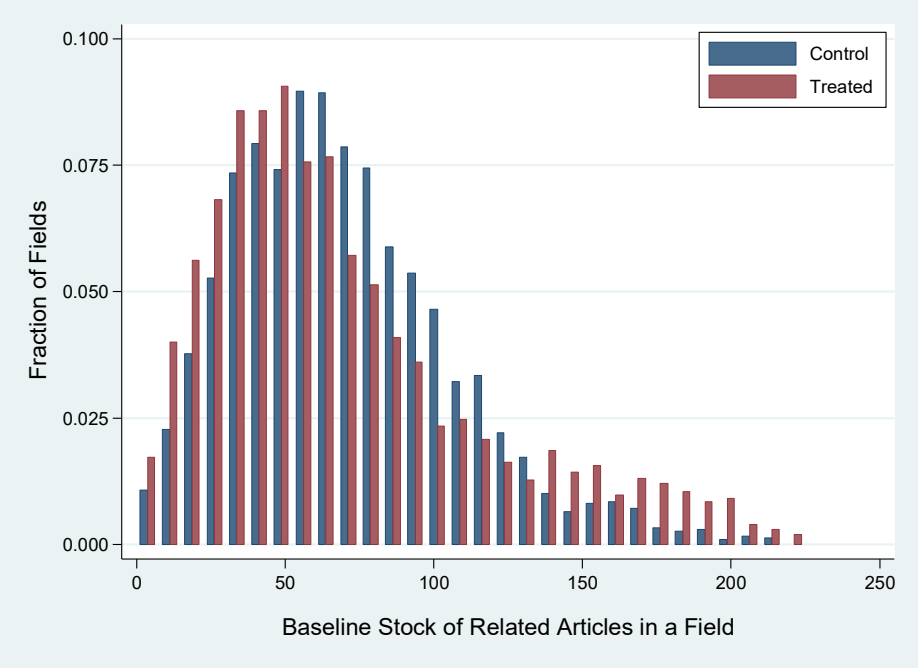


Alan Perelson (1947-)
PhD, 1972
Theoretical Biology & Biophysics, Los Alamos National Lab
In 2003: 173 pubs., 19,817 citations, $7.6 mn. in NIH funding

Ira Herskowitz (1946-2003)
PhD, 1971
Yeast genetics, UCSF
In 2003: 153 pubs., 21,549 citations, $16.8 mn. in NIH funding

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.

Figure D2: 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 3,074 treated subfields and 31,142 control subfields.

Appendix E: Robustness Checks and Extensions

Table E1: Main effect of superstar death on publication flows, aggregated up to the level of the star scientist

	All Authors	Collaborators Only	Non-Collaborators Only
After Death	0.050**	-0.191**	0.069**
	(0.004)	(0.016)	(0.004)
Nb. of Investigators	6,117	6,117	6,117
Nb. of Field-Year Obs.	263,031	263,031	263,031
Log Likelihood	-1,041,849	-273,933	-997,012

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. 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 $100\times(\exp[0.069]-1)=7.14\%$. 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 (QML) standard errors in parentheses. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$.

Table E2: Robustness Checks

	No Control Subfields			OLS Estimates		
	All Authors	Collabs. Only	Non-Collabs. Only	All Authors	Collabs. Only	Non-Collabs. Only
After Death	0.054†	0.029	0.057†	0.259**	-0.076**	0.335**
	(0.029)	(0.050)	(0.029)	(0.098)	(0.025)	(0.089)
Nb. of Investigators	452	431	452	6,260	6,260	6,260
Nb. of Fields	3,076	2,887	3,076	34,211	34,211	34,211
Nb. of Field-Year Obs.	111,705	104,791	111,705	1,258,911	1,258,911	1,258,911
Log Likelihood	-242,248	-55,558	-232,518			
Adjusted R^2				0.446	0.291	0.414

Note: Estimates stem from conditional (subfield) fixed effects Poisson specifications (columns 1, 2, and 3) or OLS specifications with subfield fixed effects (columns 4, 5, and 6). 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 E3: 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	0.105**	-0.006	0.067*	0.078*
	(0.036)	(0.038)	(0.033)	(0.033)
Nb. of Investigators	5,542	1,936	5,187	4,233
Nb. of Fields	27,015	7,196	18,079	16,132
Nb. of Field-Year Obs.	994,891	264,020	664,650	594,261
Log Likelihood	-2,050,942	-545,385	-1,339,319	-1,234,116

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, 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 (QML) standard errors in parentheses, clustered at the level of the star scientist.

$^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$.

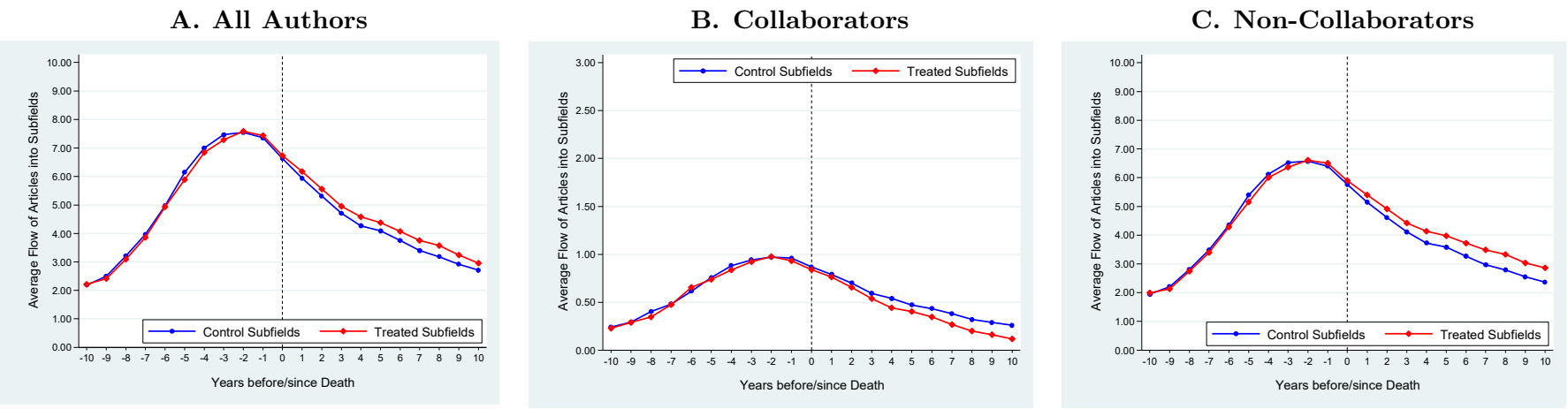
Table E4: Influence of field overlap between related authors and the stars on the rate of entry into subfields

	Below Median	Btw. 50th and 75th pctl.	Btw. 75th and 95th pctl.	Above 95th pctl.
After Death	0.128**	0.077**	0.028	0.012†
	(0.033)	(0.017)	(0.019)	(0.007)
Nb. of Investigators	6,260	6,260	6,260	6,260
Nb. of Fields	34,211	34,211	34,211	34,211
Nb. of Field-Year Obs.	1,258,911	1,258,911	1,258,911	1,258,911
Adjusted R^2	0.299	0.232	0.231	0.144

Note: This table displays some of the results depicted in Figure 2, Panel B in regression form. Estimates stem from OLS specifications with subfield fixed effects. The dependent variable is the total number of publications in a subfield in a particular year, broken into four bins: 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 (see Figure 2, Panel A); 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. 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.

$^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$.

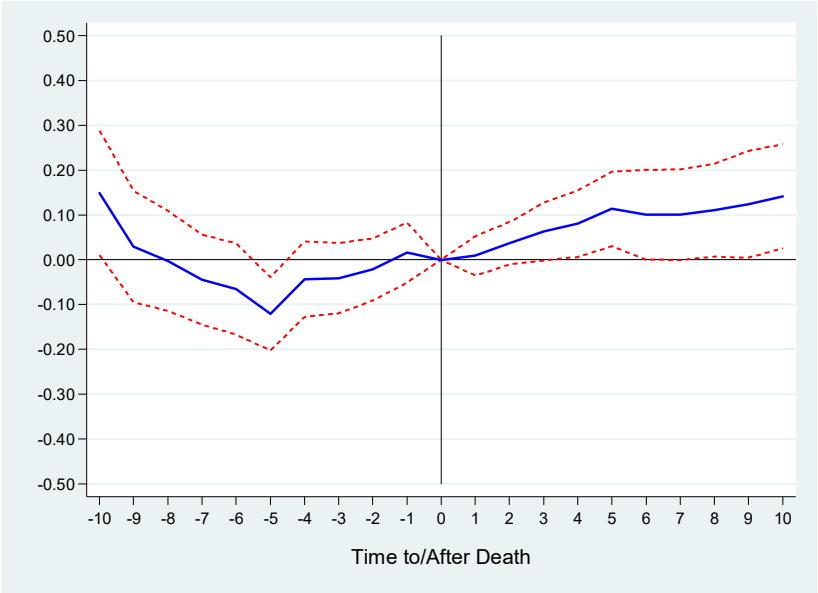
Figure E1: Effect of Star Scientist Death on Subfield Growth and Decline
Raw Data, Balanced Panel



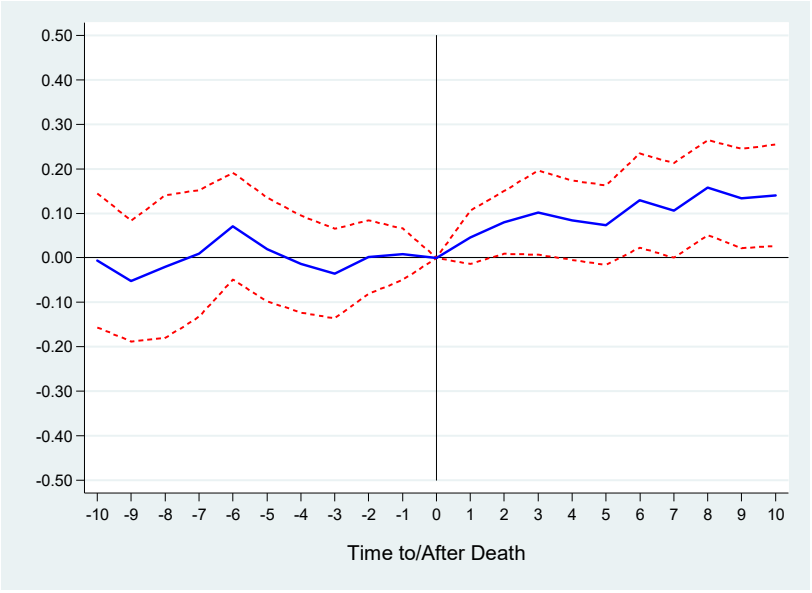
Note: Panels A, B, and C show the path of mean publication activity for treated and control subfields around the year of star death, broken down by total number of publications in the subfield (Panel A), number of publications in the subfield with a coauthor of the star (for treated and control stars alike, Panel B), and number of publications in the subfield without any coauthor of the star (Panel C). The sample used to calculate these means differs in one respect from our main sample: it has been extended from 2006 to 2012, which results in an almost perfectly balanced sample.

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

A. Anticipated Death



B. Sudden Death



Note: The graphs in this figure are patterned after Panel C in Figure 1 in the main body of the manuscript. The solid blue lines 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). Our main sample stops the observation window in 2006, since many of the covariates needed to generate the estimates in Tables 5 through 8 are not available after 2006. When the analysis is restricted to the years 1970-2006 (i.e., with an unbalanced sample), the event study graphs look substantially similar to those above. The 95% confidence interval (corresponding to (QML) robust standard errors, clustered around star scientist) around these estimates is plotted with dashed red lines.

Appendix F: Effects on Prior Agenda

Table F1: In-field vs. Out-of-field Publication Activity

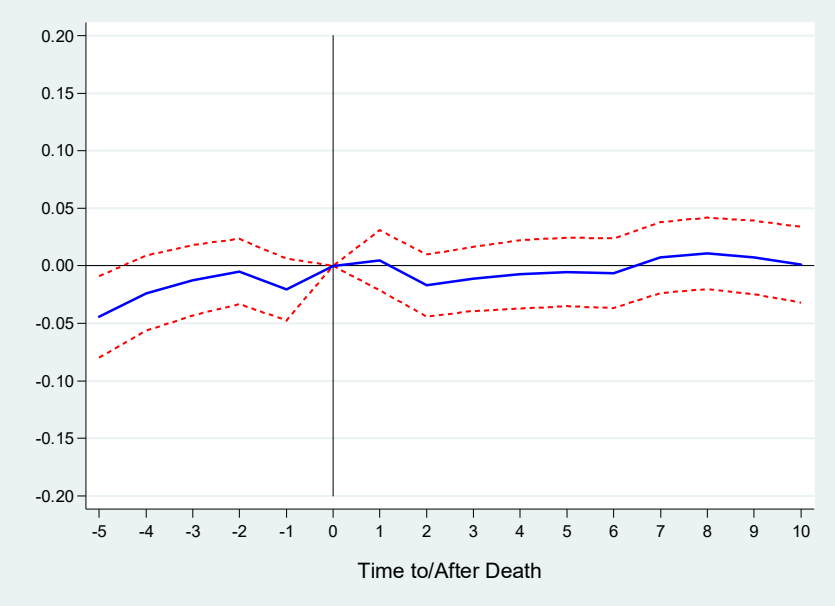
Panel A: Earliest Treating Star	In-field			Out-of-field		
	All Pubs	Middle-Authored Pubs	First/Last Authored Pubs	All Pubs	Middle-Authored Pubs	First/Last Authored Pubs
After Death	0.241**	0.235**	0.231**	0.012	0.002	0.025
	(0.035)	(0.045)	(0.051)	(0.012)	(0.012)	(0.016)
Nb. of Investigators	58,433	33,973	26,880	89,270	87,709	86,014
Nb. of Investigator-Year Obs.	1,009,696	609,185	449,465	1,361,783	1,356,063	1,342,740
Log Likelihood	-173,585	-99,792	-75,283	-2,152,402	-1,725,768	-1,484,472
Panel B: Most Related Treating Star	**In-field**			**Out-of-field**		
	All Pubs	Middle-Authored Pubs	First/Last Authored Pubs	All Pubs	Middle-Authored Pubs	First/Last Authored Pubs
After Death	0.182**	0.159**	0.182**	-0.002	-0.004	-0.012
	(0.032)	(0.041)	(0.043)	(0.014)	(0.012)	(0.018)
Nb. of Investigators	61,342	36,283	31,862	89,270	87,709	86,007
Nb. of Investigator-Year Obs.	1,067,915	654,804	558,313	1,361,783	1,356,036	1,342,716
Log Likelihood	-237,517	-140,463	-105,523	-2,144,005	-1,718,901	-1,477,885

Note: Estimates stem from conditional (author) fixed effects Poisson specifications. The dependent variable is the publication output for a related, non-collaborating author in a particular year. The first series of three columns restrict output to publications that fall in the field of the treating star. The second series of three columns restrict output to publications that fall outside of the field of the treating star. 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, to address the concern that age, year and individual fixed effects may not fully account for trends in publication output around the time the star's death. The number of observations varies slightly across columns because the conditional fixed effects specification drops observations corresponding to authors for which there is no variation in output over the entire observation period. 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 non-collaborator authors see an increase in the number of their published output in the field of the deceased star after s/he passes away—a statistically significant $100\times(\exp[0.182]-1)=19.96\%$.

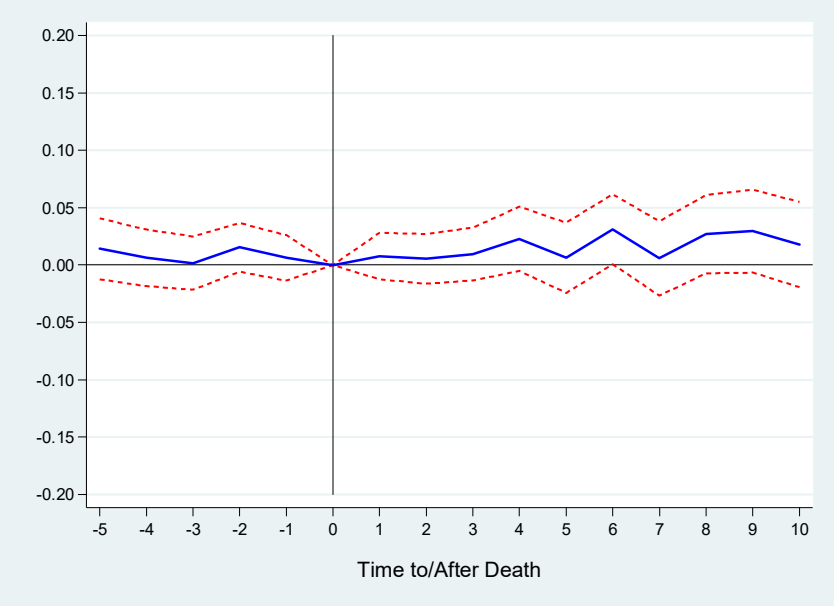
Robust (QML) standard errors in parentheses. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$.

Figure F1: Effect of Star Scientist Death on Non-collaborating Related Author Out-of-Field Publication Output

A. Earliest Star Death



B. Most Related Star



Note: The solid blue lines in the above plots correspond to coefficient estimates stemming from conditional (author) fixed effects Poisson specifications in which out-of-field publication output for a related, non-collaborating author 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 for both the treated and control subfields. The 95% confidence interval (corresponding to robust standard errors) around these estimates is plotted with dashed red lines; Panel A corresponds to a dynamic version of the specification in the fourth column of Table F1, Panel A; Panel B corresponds to a dynamic version of the specification in the fourth column of Table F1, Panel B.

Appendix G: 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	physchological 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 Kakumaga	[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

References

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.

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

Callon, Michel, Jean-Philippe Courtial, and F. Laville. 1991. "Co-word Analysis as a Tool for Describing the Network of Interactions between Basic and Technological Research: The Case of Polymer Chemistry." *Scientometrics* **22**(1): 155-205.

He, Qin. 1999. "Knowledge Discovery Through Co-Word Analysis." *Library Trends* **48**(1): 133-159.

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).

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

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.