Long-term Effects from Early Exposure to Research: Evidence from the NIH "Yellow Berets"

Pierre Azoulay[a,b,*] Wesley H. Greenblatt[a] Misty L. Heggeness[c]

[a] MIT Sloan School of Management, 100 Main Street, Cambridge, MA, 02142, United States
[b] National Bureau of Economic Research, 1050 Massachusetts Avenue, Cambridge, MA, 02138, United States
[c] U.S. Census Bureau, Research and Methodology Directorate, 4600 Silver Hill Road, Suitland, MD, 20746, United States

[*] Corresponding author.
E-mail addresses: pazoulay@mit.edu (P. Azoulay), wesley.greenblatt@mit.edu (W. Greenblatt), misty.l.heggeness@census.gov (M. Heggeness).

July 18, 2021

Abstract

Can a relatively short but intense exposure to frontier research alter the career trajectories of potential innovators? To answer this question, we study the careers and productivity of 3,075 medical school graduates who applied to the Associate Training Programs (ATP) of the National Institutes of Health (NIH) during the turbulent period of the Vietnam War, 1965-1975. Carefully selecting on observables, we compare physicians who attended the program to those who passed a first admission screen but were ultimately not selected. We find that program participants were twice as likely to choose a research-focused position after training, and considerably less likely to switch to purely clinical endeavors as their careers unfolded. Over the life cycle, NIH trainees also garnered publications, citations, and grant funding at a much higher rate than synthetic controls, and went on to mentor more trainees who themselves became successful researchers. The direction of their research efforts was durably imprinted by their training experience. In particular, NIH trainees appear to have acquired a distinct "translational" style of biomedical research which became an implicit training model for physician-scientists as ATP alumni came to occupy the commanding heights of academic medicine throughout the United States.

JEL Classification: J24, O31, O33

Keywords: Biomedical workforce, Scientific and technical human capital, Career imprinting, Mentorship, Translational medicine

This project was supported by the National Institute on Aging (grant number R24-AG048059) to the National Bureau of Economic Research. We thank Dr. Michael Gottesman, Dr. Michael Lauer, and Dr. Kay Lund for their steadfast support and encouragement. We are indebted to Dr. Anthony Fauci, Dr. Michael Gottesman, Dr. Richard G. Wyatt, and Dr. John Gallin for reflecting on their experiences in the ATP and encouraging us to investigate the impact of the NIH Intramural Program during the Vietnam War. We would also like to thank Daniel Boehlert, Andrew Breazeale, Delaney Cruickshank, Daphne Hines, Maria Isabel Larenas, Kyle Myers, Ryan Pfirrmann-Powell, Lindsey Raymond, and Ying Zeng for their diligent efforts in sorting, entering, matching, and curating the data. Lastly, this project would not be feasible without the help of Barbara Harkins and Christopher Wanjek from the NIH Office of History. We also thank Scott Stern as well as seminar audiences at Georgia Tech, Wisconsin, and Northwestern for constructive feedback. Any opinions (and all errors) are solely those of the authors and do not represent any official position of the U.S. Census Bureau or NBER.

> "[The ATP] did not help [my career], it made it... I followed a pathway that was a combination of hard work, some talent and being in the right place at the right time... None of that would have happened had I not come down here as a Clinical Associate... [I would have] gone to Vietnam for a few years in the Navy, [and then] I would have probably returned to New York Hospital. I would probably be practicing medicine right now on 69^{th} Street and First Avenue. The Clinical Associate program put me on a career track that I am still on."
>
> Anthony Fauci, Director, NIAID
> *Oral History (1998)*

1 Introduction

It has become a truism among policy-makers that innovation and technological advances are a key determinant of economic growth (Aghion and Howitt, 1992; Romer, 1990; Solow, 1957). But innovation is fundamentally constrained by the supply of innovators—those individuals whose skills and knowledge put them at risk of bringing forth a useful "new-to-the-world" idea. Innovators are made, rather than simply born, and growth possibilities are shaped by the institutions, incentives, and norms that nudge would-be innovators to receive the training necessary to bring themselves to the frontier. Indeed, over the past century, macro evidence suggests that only by steadily increasing the number of workers engaged in formal R&D activities has a steady growth rate in income per capita been sustained (Jones, 1995).

In the medium run at least, designing institutions that might increase the supply of potential innovators is therefore of crucial policy importance. Yet, severe headwinds frustrate efforts to broaden the innovator pipeline. First, because scientific and engineering training is protracted, individual career choices are often shrouded in uncertainty, both with respect to the monetary payoffs and the direction of human capital investments likely to earn the best labor market returns. Witness, for example, the dismal track record of "manpower analysis" and the perennially flawed predictions of "innovator shortage" (Freeman, 1975; Teitelbaum, 2014). Second, innovative careers are fragile (Milojevic et al., 2018) both because of the winner-take-most aspect of the scientific reward system, and because skills at the frontier depreciate rapidly, leading many initial entrants to abandon the idea sector and reenter the production sector (Deming and Noray, 2018). Third, especially for countries with domestic training capabilities, restrictions on high-skilled immigration can act as a brake on plugging leaks in the innovator pipeline (Kerr, 2018). As a result of these headwinds and the elimination of mandatory retirement in academia in the mid-1990s, the scientific workforce is aging rapidly (Blau and Weinberg, 2017).

Despite the paucity of research examining the allocation of talent to innovative activities, some recent evidence points to an important friction, that of exposure to research during an individual's formative years. In a telling anecdote, pro-footballer turned Math PhD student John Urschel recounts how his athletic prowess was identified and nurtured from a young age, whereas his mathematical talents were left undeveloped until a chance encounter with an inquisitive college instructor (Urschel and Thomas, 2019). More systematically, Bell et al. (2019), using IRS tax records linked to U.S. patent data, provide evidence of a strong association between fathers and sons' propensity to patent in the exact same narrow patent class, a finding most easily explained by early socialization opportunities regarding the feasibility and desirability of a research career.

The existence of exposure effects might at first blush appear surprising, but their potential importance is better appreciated if one remembers that early research careers exhibit both brittleness—in the sense that small negative shocks can shift individuals back to the production sector of the economy (Hill, 2018)—and malleability—in the sense that the flexibility to alter one's research trajectory declines over the life cycle (Higgins, 2005). Together, brittleness and malleability suggest that transient but intense formative experiences in the early career may significantly influence potential innovators' decision to enter the "ideas sector" of the economy, as well as their choice of research trajectory, domain, or methodology.

Despite the empirical plausibility of exposure effects, providing convincing evidence of their existence and magnitude presents seemingly insurmountable challenges. Three necessary ingredients are required. First, one needs to identify a population of "naïve to research" individuals who nonetheless possess much of the human capital required to propel themselves to the research frontier. Second, one requires an intervention consisting of a short but intense exposure to research in a rarefied intellectual environment to a (preferably random) subset of this population. A final requirement is the opportunity to observe these individuals for a long period with minimal loss to follow-up, and see their career unfold.

In this paper, we study an intervention in physician training that comes close to bringing together these three ingredients: The Associate Training Program (ATP) of the National Institutes of Health (NIH). The ATP brought recent MD graduates to the intramural campus of the NIH in Bethesda, Maryland for two to three years to participate in research under the supervision of NIH investigators. A unique aspect of the program is that participation fulfilled a draftee's military service requirement (Berry, 1976). After the war ended, trainees began to refer to themselves ironically as "Yellow Berets," a derogatory term used to contrast draft dodgers with the elite Green Berets—the U.S. Army Special Forces (Baskir and Strauss, 1978; Klein, 1998). Though quite small when the program was founded in 1953, its scale steadily grew with applications dramatically increasing during the years of the Vietnam War. The ATP can be considered a large human capital intervention not because it selected a particularly large cohort (even at its 1973 peak, the program drafted only 229 associates, or approximately 2.5% of graduating male students) but because it induced a very high proportion of eligible participants to actually apply, from around 20% in 1963 to close to 80% in 1971.[1] Though some applicants had prior exposure to biomedical research in medical school or during their undergraduate studies, the unpopularity of the war drove many physicians who otherwise would not have been interested in a research career to apply for one of those coveted positions (Varmus, 2009). This unique confluence of events provides us with a quasi-experimental lever to disentangle the role of sorting from that of training and mentorship, always a vexing challenge in empirical studies of the scientific labor market.

We study the careers and productivity of all 3,075 male medical school graduates who applied to the ATP and were interviewed on campus between 1965 and 1975. We build a rich hand-collected data set containing the complete training and career histories for these individuals, including all publications, patents, NIH grants, and citations. Carefully selecting on observables, we compare physicians who attended the program to those who passed a first admission screen but were ultimately not selected. Despite lasting only two to three years, we find that the ATP had a large and sustained impact on the careers of those who attended. Relative to synthetic control applicants, program participants were twice as likely to sort into research-focused positions, and dramatically less prone to switch to purely clinical endeavors as their careers unfolded. Over the life cycle, NIH trainees also garnered publications, citations, and grant funding at a much higher rate than synthetic controls, with over a 75% higher odds of joining the biomedical research elite.[2] They also mentored more trainees who themselves became successful researchers, providing a way their impact could persist through the training of the next generation. Moreover, the direction of their research efforts was durably imprinted by their training experience. In particular, ATP attendees appear to have acquired a distinct "translational" style of biomedical research

[1] Since records on the total number of applicants in each year have not survived, the first figure comes from a back of the envelope calculation (see footnote 4), whereas the second stems from anecdotal accounts that are plausible, but hard to substantiate empirically.

[2] Defined as receiving the Nobel Prize, being appointed Howard Hughes Medical Institute investigator, being elected to the National Academy of Science/Medicine, or winning an NIH R37 MERIT award.

which became an implicit training model for physician-scientists as ATP alumni came to occupy the commanding heights of academic medicine throughout the United States (Khot et al., 2011).

In addition to the unique historical importance of the NIH ATP (Klein, 1998), our study sheds light on the forces that shape skill acquisition in medicine, and how medical training influences the rate and direction of medical progress. Much of the training physicians receive in medical school, internship, and residency is fungible between medical care and medical research. Early in their career, physicians invest heavily in human capital, but then typically go on to apply their skills narrowly, for the benefits of their (private) patients. These same skills, however, can be redeployed in research activities, where physician effort also generates social returns. In fact, it has been a long-standing policy goal of the medical elite to steer a larger number of physicians towards research careers (Wyngaarden, 1979). As a result, studying the NIH training programs in the Vietnam War era provides a unique window on the long-term consequences of exogenously shifting a well-defined population from the "production sector" of the economy (i.e., clinical care) to its "ideas sector" (i.e., biomedical research, including bench, clinical, and translational research).

Our study also speaks to how the institutional environment of scientific training programs shapes their participants' research careers. A limited number of studies have examined how mentorship during or after training (Ginther et al., 2020; Shibayama, 2019), funding level and source (Blume-Kohout and Adhikari, 2016; Broström, 2019; Ginther and Heggeness, 2020; Jacob and Lefgren, 2011), and their interaction with trainee background (Graddy-Reed et al., 2019) may impact the outcomes of pre- and post-doctoral scientific training programs. By studying medical doctors who were pushed to seek research training by a unique confluence of historical events but many of whom ultimately received training in other settings, the NIH ATP provides a lens on how the content of training programs shapes the type and quality of the scientific talent it nurtures.

The rest of the manuscript proceeds as follows. Section 2 provides institutional background on the NIH ATP program, including the procedures used to select the trainees. Section 3 describes our sample construction and provides descriptive statistics. Section 4 discusses our econometric approach, while Section 5 presents our main results. Section 6 puts the results in context, and discusses their implications for the design of scientific training programs in the twenty-first century.

2 Institutional Setting

Relative to other professional or creative endeavors, the scientific labor market is notable for the extent to which, at any given point of time, a handful of research institutions are responsible for training a disproportionate share of the future elite in a field while simultaneously providing an extraordinary environment for breakthrough discoveries. Examples abound from a wide variety of scientific fields. In physics, the Cavendish laboratory was the prime breeding ground of atomic physicists in the first half of the twentieth century (Rhodes, 1986); the Laboratory of Molecular Biology, also located at the University of Cambridge, played a similar role for biomedical research after the second world war (Bynum, 2012; Rubin, 2006). This phenomenon is not limited to the physical sciences. For example, the MIT economics department stands out from those located at other universities in the extent to which it spawned a community of academics who went on to exert a profound influence on the discipline (Svorenčík, 2014).

During the period of our study, the intramural campus of the NIH, located in Bethesda, Maryland, was widely recognized as one of the preeminent biomedical research institutions. One aspect setting it apart from other elite institutions, however, was its unique ability to attract recently minted physicians eager to pursue a research career. Due to the confluence of multiple factors—the Doctor Draft, plentiful federal funding, and the opening of a massive clinical research center in 1953—the NIH had probably no equal in the world with respect to the training of "physician-scientists" (Park, 2003). We draw on historical evidence, including a large archive of oral histories curated by the NIH Office of History to describe this setting in more detail, review the genesis and development of the Associate Training Program (ATP), and describe how trainees were selected and trained during this period (see Appendix E for additional details).

2.1 The Associate Training Program

The NIH ATP started in 1953 with about 15 medical graduates to provide research training to physicians (Klein, 1998). Associates would come to Bethesda and do research under the supervision of NIH investigators, usually after completing a portion of their residency training. Two years were typically spent in the program, with the option to extend training an additional year. From the start, the program was focused on turning physicians into independent medical investigators well-grounded in scientific knowledge and methods. The goal was on learning how to do research more than simply doing research itself and on bringing the physicians into close contact with accomplished scientists. In addition to the research, the NIH also hosted a set of after-hours basic sciences courses for program participants that could rival the offerings of major universities. Christian Anfinsen, a Nobel Laureate and NIH investigator during the early years of the program, describes its key features as *"... the importance of having the [associates] work on problems of [their] own choice rather than be 'servants' in the research problems of the preceptor, and the importance of providing the student[s] with some integrated and organized basic knowledge as a foundation that would permit them to do their own integrating of knowledge later"* (Anfinsen, 1963). While the focus was on research, some participants were able to get credit for their time at the NIH towards their required clinical training for board certification.

By the early 1960s, the Associate Training Program had been expanded to include three separate tracks. Clinical associates would divide their time between clinical care at the NIH Clinical Center and laboratory research. Research associates would spend most of their time on research and had limited clinical responsibilities. Staff associates also had training in research administration as well as undertaking clinical or laboratory research.

Oral histories from NIH staff are replete with claims attesting to the cutting edge research, breadth of expertise, and concentration of talent in biomedical research within the confines of the intramural campus that resulted in a rarefied environment (Appendix E). In addition, many ATP fellows came to view the focus on what would later be called translational research as a distinctive element of the approach to research at the NIH. This was no accident. James Shannon, one of the early leaders of the NIH, carefully structured the intramural program to facilitate close cooperation between basic and clinical research (Goldstein and Brown, 1997; Park, 2003). Anthony Fauci, an ATP alumni and prominent HIV/AIDS researcher, recalls, *"What the Clinical Associate Program does is it gives you a very interesting perspective on the relationship between disease and the basic science that you have to study to be able to approach disease... Also the link, as we used to say, between 'the bed and the bench,' you see something at the bedside, you bring it back and ask the question at the bench or you make a discovery at the bench and you go back and apply it to the bedside, that bedside to bench phenomena was really what the Clinical Associates program was all about"* (Fauci, 1998).

Since the NIH, through historical accident, grew out of a laboratory within one of the U.S. Navy Marine Hospitals, ATP applicants applied to the program under the auspices of the U.S. Public Health Service and those selected became commissioned officers. This allowed service with the U.S. Public Health Service to fulfill any military service obligation a physician may have if drafted.[3] The interest in and level of competition for spots in the program increased in proportion to the perceived hardship of military service. The program, however, was highly competitive even before the increased interest during the Vietnam War. Unfortunately, there is no reliable information on the total number of applicants to the program, except in a single year before the start of our information period: 1963. That year 53 of 1,464 physician applicants were selected (NIH Office of Research Information, 1963).[4] At its peak, in 1973, the program included 229 associates (Klein, 1998). In contrast, in the year following the 1973 Paris Peace Accords which effectively led to an end to the military draft, the NIH was not able to fill its associateship quota for the year, and by 1976 included only 108 physicians, down over 50% from its peak (Klein, 1998).

While certainly some of the physicians would have applied to and attended the program regardless of the war, avoiding the draft was a significant motivation. Donald Fredrickson, a former director of the NIH and one of the first clinical associates in the program in 1953, later played a role in determining who to admit to the program during the 1960s and 1970s. He recalled, *"The NIH Associates program would never have been as popular or as competitive as it was without the draft"* (Fredrickson, 1998). Anthony Fauci, a program alumni and Director of the National Institute of Allergy and Infectious Disease, echoed these sentiments *"...every single physician went into military service...essentially, I came down to the NIH because I didn't have any choice"* (Fauci, 1989).

2.2 The Application Process

Applications to the NIH ATP were typically submitted two years in advance, during the final year of medical school with a planned program start date after completing internship and the first year of residency training. Applications included academic transcripts, references, publications, and planned post-graduate training institutions. After a first screen based on these documents, a small number of applicants were invited to interview on campus at the NIH in order to match with a particular laboratory and mentor. Unfortunately, much of this written documentation was destroyed, leaving only the application index cards of the subset of candidates who cleared the first admission hurdle and attempted to match with a laboratory. There is also no official record of the labs with which each participant

[3] Of note, in addition to the NIH, the U.S. Public Health Service had other programs through which physicians could apply to spend two years of service, including at the Center for Disease Control, the Food and Drug Administration, and the Indian Health Service.

[4] In 1963 there were 7,265 graduates from US Medical Schools (Association of American Medical Colleges, 2016), an estimated 5.6% of which were female (Snyder, 1993). Using this, we can conclude approximately 21% of eligible male medical students actually applied to the NIH ATP in 1963.

attempted to match or offers made. The data can only tell us that out of these second round applicants, roughly 63% accepted an ATP position and attended the program. According to the NIH's official documentation, these final appointments were made based upon intellectual attainment and demonstrated research interest and ability (NIH, 1968).

Applicants were undoubtedly positively selected from the eligible population—male medical school graduates. In Appendix C, Table C1, we can see that compared to a random sample of non-applicants drawn from the American Medical Association (AMA) Physician Master File, applicants graduated from more selective medical schools (as measured by NIH grants) and published at significantly higher rates than non-applicants before application (0.9 vs. 0.3 publications on average). However, it would be wrong to conclude from this evidence that applicants displayed a preternatural disposition for research career prior to application. For instance, the median number of publications for applicants is zero; the overwhelming majority of applicants do not hold a PhD degree; and applicants do not appear particularly precocious, relative to the eligible population (kernel densities corresponding to the age distribution at the time of application for applicants and non applicants is depicted in Figure C1; the two curves are nearly identical).[5]

The oral and written historical records also speak to the difficulty in evaluating research potential and making decisions between candidates. Donald Fredrickson, an ATP alumnus who later served on the selection committee for the program in the 1960s and 1970s, recalls that *"...the main objective was getting people who would use this environment to turn into scientists,"* but also notes selecting participants was *"extremely difficult because all we really had was the scholastic record of most people. Very few had done any research...so the art of picking out of a whole group of qualified people those who might become successful scientists was extremely difficult... We would have to pick them with a certain amount of variety because our programs needed people of diverse interests"* (Fredrickson, 1998). Harry Kimball, another alumnus of the program who was also later involved in applicant selection remembers *"It was truly astonishing how qualified these people were and the kind of close decisions you had to make as to who to offer a spot in the program"* (Kimball, 1997). Harold Varmus describes how the decisive factor in his own selection into the program likely did not hinge on his promise as a budding scientist. Rather, he writes that during his interview with

[5] An additional piece of evidence argues against viewing the applicant population as being dominated by science "geniuses": matching carefully the applicant roster with the Directory of Rhodes Scholars, we found only seven matches (four treated physicians and three control physicians). Note that comparisons with "non-applicants" are subject to an important caveat: since we do not know the identity of the first-round applicants, our sample of non-applicants could in fact include individuals who did not pass the first application screen.

Ira Pastan *"My schooling in literature turned out to be more important than my interest in endocrinology, Ira's field, because Ira's wife Linda, a poet, had often complained that Ira's colleagues seldom talked about books. Ira, himself an enthusiastic reader, thought it might be helpful to have someone with my background in his lab"* (Varmus, 2009).

2.3 Prior Evaluations

A handful of prior studies have examined the program. Klein (1998) provides a thorough description of the ATP and the NIH during the Vietnam era grounded primarily in the conduct and review of historical documents and interviews. We have drawn on her analysis to provide much of the necessary institutional background required to guide our empirical analysis. Khot et al. (2011) analyze the careers of NIH ATP attendees from 1955 to 1973, comparing them to a random sample of medical school faculty that graduated in the same years selected from the Association of American Medical Colleges Faculty Roster. The authors show that relative to these controls, ATP participants were 150% more likely to achieve the rank of full professor, twice as likely to become a department chair, and three times as likely to become a medical school dean. Matching the population of attendees with a series of prestige markers appropriate for biomedical researchers, they found in their sample nine winners of the Nobel Prize in Physiology or Medicine, ten recipients of the National Medal of Science, 44 members of the National Academy of Sciences, and 125 members of the Institute of Medicine. Our study improves on their design with a more appropriate control group, that of unsuccessful applicants to the ATP, which helps shed light not simply on the effect of ATP attendance on the intensive margin—articles, citations, grants, patents—but also on the extensive margin: how did selection shape applicants choice of career, in particular participation in research activities as opposed to purely clinical endeavors?

3 Empirical Design, Data, and Descriptive Statistics

3.1 Data

The application index cards for the NIH Associate Training Programs form the raw material for the creation of our dataset. While the cards for successful applicants had been previously digitized and used in prior research efforts (e.g., Khot et al., 2011), the index card for applicants who did not attend the program were previously thought to have been destroyed. In 2015, carton boxes containing a subset of these index cards—those corresponding to applicants who interviewed on campus but were ultimately not offered a position—were discovered at the National Archives by the NIH archivist, Barbara Harkins. Figure 1 displays the number of index cards in our dataset in each year belonging to our observation window, 1965 and 1975. While the ratio of successful to unsuccessful applicants is approximately 2:1 over the entire period, this average masks large swings, with the years 1970, 1971, and 1972 exhibiting a greater proportion of unsuccessful applicants. These years correspond to the height of the Vietnam War mobilization effort.

We limited our analysis to those who applied to the program between 1965 and 1975. To arrive at the final list of 3,075 applicants, we eliminated 22 applicants who did not hold an MD degree, three unsuccessful applicants who applied at the very start of medical school (and did not reapply), and eight who died while in training, or soon thereafter. We also excluded 33 female and 22 foreign medical school graduates as their motivations to apply may have been very different from applicants subject to the draft. Despite our best effort, we also lost 13 applicants to follow-up (less than 0.42% of the total). In the case of repeated applications for the same applicant, we retained only the latest one.

For each of these physicians, we manually collected their training and career history using a mix of Google, Doximity, and LinkedIn searches; medical licensure records; professional profiles and CVs; Who's Who profiles; and other publicly available internet sources. These were supplemented with physician biographical information contained in the AMA Physician Masterfile. To ascertain treatment status, participation in the ATP was verified with the biographical resources above as well as NIH telephone directories and internal human resource records (additional details on data set construction are available in Appendix A).[6] Applicants who were appointed to the Public Health Service Commissioned Corps but served at the Center for Disease Control (CDC) or the Indian Health Service (IHS) were assigned to the control group. Of course, many members of the control group received research training in traditional academic medical settings, some of them after a period of military service, though only one applicant in the sample appears to have served in the Vietnam military theater. The final sample contains the records of 3,075 physicians (1,929 program attendees and 1,146 non-attendee controls).

[6] Our set of treated applicants include fellows who completed their training outside of the confines of the NIH intramural campus in Bethesda, such as the Baltimore Cancer Research Center or the Food and Drug Administration (FDA). Other NIH locations were even more far-flung such as the Rocky Mountain Laboratory (located in Hamilton, Montana) or the Panama Control Zone. As a robustness check, we repeated our analysis excluding the 267 ATP attendees not located on the main NIH campus in Bethesda with similar results obtained.

We distinguish between three career phases for all applicants. First, the education, or pre-application phase, which ends at the end of medical school. Second, the training phase, which covers internship, residency, post-residency fellowships, as well as national service regardless of the setting where it was served (Army/Navy, NIH, CDC, IHS). Finally, the independent phase of the career begins immediately after the end of the training phase, and ends with retirement or death. When referring to career choice in the rest of the paper, we refer to the choice of employment in this last career phase. 277 (9.01%) applicants pass away prior to their retirement; 762 (24.78%) retire prior to 2017, the end of our observation period; and for 2,036 applicants (66.21%), the career is still ongoing as of 2017. Though these observations are technically censored, it is important to acknowledge that the youngest applicant in our sample was 65 years old in 2017 and in his thirty-first career year. To a first order of approximation, these physicians are therefore at the twilight of the active phase of their research or clinical careers.

Publications, citations, patents, and NIH grants were collected for each individual from PubMed, the Web of Science, the U.S. Patent and Trademark Office (USPTO), and the NIH's Consolidated Grant Applicant File, respectively, and carefully name-disambiguated. For publications we include only original research articles, excluding other types of publications such as letters, editorials, and review articles. Importantly for our analysis, we use the richness of the individual profiles collected to measure participation in research independently of the applicants' employers. For instance, the career of many of our applicants unfolds within academic medical centers in purely clinical positions where there is no expectation of publication. In contrast, other applicants work in industry or other non-academic institutions and yet amass a respectable publication record in the context of non-traditional research careers. Since our motivation is to understand how early career interventions might influence long-run engagement with the idea sector of the economy, distinguishing between career *locus* (academic versus non-academic jobs) and career *focus* (research jobs versus clinical jobs) is important.

3.2 Descriptive statistics

Pre-application characteristics. Table 1a presents descriptive statistics regarding ATP applicants at the time of application. Applicants with stronger academic credentials, or with evidence of involvement in research activities are also more likely to attend the program. For instance, applicants holding a PhD degree, those with a publication record, those inducted in an elite medical school honors society ($A\Omega A$),[7] and those having graduated from elite medical schools (as proxied by the NIH funding received by its affiliated faculty members) are more likely to be selected.[8] Recall that these applicants all survived a first screen, so one might have expected that covariates observable before this initial screen would not influence the selection decision at the interview stage. The fact that observable markers of "research preparedness" do in fact predict selection imply that interviewing "skills" are correlated with these markers, or alternatively, that the ultimate decision makers place positive weights on them even at the second stage of the process. However, one must remember that due to the young age of the applicants, the signals of research potential upon which the selection decision relies are necessarily noisy. For instance, 59.4% of applicants have no publication to their name within two years of their ATP application (67.4% for attendees; 54.7% for non-attendees). ATP attendees also applied to more NIH institutes (3.9 vs. 2.9), perhaps signaling greater interest in or motivation for research undertakings.[9]

Career choice. Table 1b provides basic statistics regarding career outcomes, with a particular focus on the first job following the end of the training phase and the last job held by each applicant before the earliest of 2017, retirement, or death (Appendix Tables B2 and B3 provide a finer-grained occupational breakdown). It is immediately apparent that ATP attendees choose academic (76% vs. 57%) and research (69% vs. 46%) careers at a more pronounced rate, relative to non-attendees, following the end of their training. These differences reflect in part time spent in training, though this contrast is not especially stark: On average, ATP attendees spend an additional 6.7 months in post-graduate training prior to achieving career independence, relative to non-attendees. The gap does not seem to narrow as their career unfolds, though one can observe attrition in the subsample of attendees. The proportion of fellows in research positions falls from 69% to 52% between the beginning and the end of the career. Overall, these univariate comparisons corroborate the claims made by ATP alumni regarding the effect of their training on career orientation. For instance, Harry Keiser, an ATP alumnus and later clinical director of the National Heart, Lung and Blood Institute, mentions that *"if I had gone back to Northwestern. . . I would have almost certainly*

[7] Criteria for selection into $A\Omega A$ varies by school, but typically weighs academic and clinical excellence most heavily.

[8] Appendix Table B1 lists the 10 most frequent medical schools from which physicians in the sample graduated, separately for attendees and non-attendee controls. Appendix Figure B1 provides a histogram for the distribution of the number of original publications published up to the year of ATP application, weighted by the journal impact factor of the publication outlet in which they appeared.

[9] The right-most columns of Table 1a provide a comparison of means in the reweighted sample which reflects the methodology presented in Section 3.3 and Appendix F. In the pseudo-population of trainees created by this procedure, the differences in baseline covariates are no longer statistically significant.

gone out into private practice... I certainly would not have continued to devote the rest of my life to research" (Keiser, 1998).

Research outcomes. Tables 1c reports descriptive statistics on a variety of research outcomes. ATP attendees garner over twice the number of career publications on average (77.8 vs. 37.3). Similar differences can be observed for patents (1.7 vs. 0.7), NIH extramural grant funding ($12.4 vs. $4.5 million), and citation impact (5,131 vs. 1,988 for article-to-article citations; 20.2 vs. 7.5 for patent-to-patent citations). ATP attendees' publications are also more heavily cited in patents (252 vs. 80). Attendees receive greater NIH R01 funding as well, with $3.1 million compared to $1.2 million over their career.

We also examine the "fecundity" of ATP applicants by identifying the set of individuals they train over their career who go on to be awarded NIH R01 funding, a key marker of research independence in U.S. academic medicine. In this way, the impact of training institutions can ripple through a much larger community of scholars as yesterday's trainees become the trainers of today. In the context of our data, a trainee is an individual who, in a window centered on the time of her highest degree, appears as first author on a publication jointly with the ATP applicant in last authorship position. We then match the names of these individuals with the NIH Consolidated Grant Applicant File, allowing us to identify the subset of trainees who go on to be awarded NIH funding (more details are provided in Appendix H). This is a relatively sparse outcome, but there again, successful applicants appear more prolific than unsuccessful ones (0.76 vs. 0.21 R01-funded trainees and $4.7 vs. $1.2 million in trainee career R01 grants on average).

Panels A, B, and C of Appendix Figure B2 display histograms for the distribution of career publications, citations, and NIH funding by treatment status. The differences in achievement between attendees and non-attendees are even more pronounced in the right-tails of these distributions. This is also reflected in the rate at which attendees accrue markers of research excellence over the career, relative to non-attendees (Table 1d). In the control group, no physician ever receives a Nobel Prize or a Howard Hughes Medical Institute (HHMI) Investigatorship (the corresponding numbers in the treatment group are 7 and 32, respectively). The differences in the rate at which treatment and control physicians become Members of the National Academies or NIH MERIT awardees are less stark, but still large in magnitude.

Research style. We develop a battery of measures to capture differences in research style across physicians in the sample. In particular, we take a first stab at measuring "translational" biomedical research. Translational research does not have an agreed-upon definition (Butler, 2008; Woolf, 2008). For the purposes of this paper, we will build upon the view of David Nathan, an NIH ATP alumni and former president of the Dana-Farber Cancer Institute (2005):

> *"Translational clinical investigators come in at least two flavors... One class includes physician-scientists interested in disease mechanisms... But these almost never interact in their research with an intact patient/subject. Such disease-oriented researchers are content to study tissue samples, cell lines, and model systems such as mice, fish, and yeast and do so with great benefit... Their career paths are only slightly distinguishable from those of basic scientists... The other class of physician-scientists include patient oriented researchers. They actively search for patients who may enable them to uncover the secrets of complex diseases, care for those patients, and with their permission, undertake to explore new diagnostic and therapeutic approaches to treating their diseases."*

As a concrete (and famous) example of translational research of the first type, consider the work of NIH ATP alumni Joseph Goldstein and Michael Brown, recipients of the 1985 Nobel Prize for Medicine and Physiology. Their initial investigations were inspired by observations of patients with familial hypercholesterolemia they saw at the NIH Clinical Center (Goldstein and Brown, 1997). Through patient-inspired basic investigations performed at the laboratory bench, they identified the underlying root cause of this disease as a lack of low-density lipoprotein receptors. These discoveries in turn informed drug development efforts, ultimately leading to the market introduction of statins. The work of Goldstein and Brown illustrates well the importance of both the "bench to bedside" and "bedside to bench" transitions which are a recurring theme in the oral histories of ATP alumni.

Conversely, Philip Pizzo personifies an approach to translational research closely connected with patient care. After his clinical associateship, Pizzo stayed on at NIH, becoming Chief of Pediatrics and Scientific Director of the Division of Clinical Sciences at the National Cancer Institute before being named Physician-in-Chief of Boston Children's Hospital and later Dean of Stanford Medical School. An expert in infectious disease and cancer, examples of his contributions include the first use of antiretroviral medication in children with HIV, a phase I trial of a solubilized receptor used by HIV for cell attachment, assessing the effectiveness in cancer patients of a diagnostic test for invasive fungal infection previously studied only in animal models, and in vitro testing of approaches to rescue neutrophil dysfunction using HIV patient samples.

The MeSH thesaurus from the National Library of Medicine provides the raw material necessary to create our measures of research style. MeSH consists of terms arranged in a hierarchical structure that permit searching at various levels of specificity (there are over 29,000 descriptors in the 2019 edition of MeSH). Almost every publication in *PubMed* is tagged with a set of MeSH terms (between 1 and 68 in the current edition of *PubMed*, with both the mean and median approximately equal to 10). For each article published by a scientist in the sample, we measure disease orientation by the presence of a disease MeSH term. To capture bench research, we take note of the presence of MeSH terms for molecular biology techniques—such as *nucleic acid amplification techniques* or *cell migration assays*, MeSH terms corresponding to model organisms—such as the nematode *caenorhabditis elegans* or the fruit fly *drosophila melanogaster*, MeSH terms related to cellular structures and macromolecules—e.g., *DNA topoisomerase IV*, or MeSH terms denoting biochemical and cellular processes—e.g., *oxidative phosphorylation* (See Appendix G for further details).

In a second step, we partition the bibliome into four mutually exclusive styles: (i) *Basic science* articles are not disease-oriented, are tagged by at least one bench science keyword, and are not clinical trials; (ii) *translational* articles are disease-oriented, tagged by at least one bench science keyword, and not clinical trials; (iii) clinical trials (identified using MeSH terms and the publication type field in *PubMed*); and (iv) *"other" clinical* articles, which are disease-oriented, not clinical trials, and not tagged by any bench MeSH keywords.[10]

We create four additional approaches to uncover the empirical signature of a translational research style. First, a natural way for the transition from bench to bedside to take place is for clinical researchers to further develop translational work, for example by performing a clinical trial. We designate an article as "inspiring translational research" whenever it is translational according to the above criteria and is cited by a clinical trial publication. Second, in the same spirit, we identify work that "builds on translational research": articles that report the results of a clinical trial and also list a translational publication in their references. Third, we identify papers published in six high-impact journals that prominently advertise their translational focus (the *Journal of Clinical Investigation*, the *Journal of Translational Medicine*, *Science Translational Medicine*, *Nature Medicine*, *Translational Research: The Journal of Laboratory and Clinical Medicine*, and the *Journal of Experimental Medicine*). Finally, a different way to facilitate the bench-to-bedside transition is to enable biopharmaceutical firms to build on the applicants's published research, since many health-related innovations cannot reach patients unless firms invest in bringing them to market (Azoulay et al., 2009). To capture this, we

[10] Jointly, these styles comprise 93% of the applicant's published output. For the style analysis, we ignore the residual unclassifiable publications.

tag each article that garners at least one citation in the header of a patent subsequently granted by the USPTO (Marx and Fuegi, 2020). This provides a crude way to capture the extent to which biopharmaceutical firms build on the work of the scientists in the sample to inform their applied R&D efforts.

Table 1e reports descriptive statistics for the research style measures. Because these measures are only meaningfully defined for publishing researchers, we create a subsample that only includes the 2,584 scientists (1,730 treated and 854 controls) who publish at least one article after the end of their training. Rather than focusing on the levels of these variables, we normalize them by the total number of articles published by each scientist in the independent career phase.

Non-attendees and attendees differ markedly in the style composition of their published work. The proportion of basic science articles is almost twice as high for successful applicants (19.9% vs. 10.7%); the proportion of translational articles is approximately 30% higher; and the proportion of clinical trials is approximately 10% higher. This means that a higher fraction of the non-attendees' output falls into the "other" clinical category. Similarly, univariate comparisons point to higher translational orientation for attendees, relative to non-attendees, using additional measures of research style. For instance, a higher fraction of attendees' articles appear in a small set of explicitly translational journals, are referenced in patents, or inspire follow-on translational research. Below, we explore whether these differences subsist when comparing treated and control physicians with similar observable characteristics.

3.3 Econometric Considerations

The univariate comparisons point to large differences in outcomes between attendees and non-attendees of the NIH ATP. It would be hazardous to interpret these differences as reflecting the causal effect of the ATP "treatment," since it is obviously a goal of NIH laboratory heads to admit applicants with the most research promise. Recall that all applicants in our sample already passed a first selection screen. Yet residual sources of selection might remain at the interview stage, e.g., the admissions committee might extract relevant information regarding an applicant's suitability for a research career in a series of relatively short interviews. To address this fundamental identification challenge, we adopt a propensity score weighting methodology which belongs to a broad class of "selection-on-observables" techniques (additional details are provided in Appendix F).

Inverse probability of treatment weighted estimation. Let us assume that the NIH principal investigators recruiting fellows at the interview stage are unable to select applicants on the basis of covariates unobserved by the econometrician and correlated with research career success—the "unconfoundedness" assumption. This assumption is not refutable and it places strong demands on the data generating process. In addition, we must assume that, for all included values of the covariates predicting treatment, the likelihood of being selected to attend is positive—the "common support" assumption. Under these assumptions, Hirano and Imbens (2001) show that various treatment effects of attending the NIH ATP, conditional on exogenous applicant characteristics, can be recovered by weighted least squares or weighted maximum likelihood estimation where the weights correspond to the inverse probability that each observation is treated. Our weighting procedure effectively creates a pseudo-population of applicants in which observable covariates no longer predict assignment to treatment and the causal association between treatment and the outcome variable is unchanged from the original population. We refer to this as the Inverse Probability of Treatment Weighted (IPTW) estimation (Austin and Stuart, 2015; Xu et al., 2010).

Informative censoring. Although we focused on the problem of non-random selection into treatment, a second problem arises because some applicants might fail to engage in research activities for the sole reason that their chosen position does not afford them the possibility to publish, seek external grants, or train the next generation of scientists. This problem is distinct from informative loss to follow-up. These physicians' careers are observed in full and yet it does not seem meaningful to compare the research productivity of a full-time, tenure-track academic researcher with that of a clinician who very occasionally dabbles in research. We deal with this problem by treating early exit from research as another treatment. As Robins et al. (2000) note, adjusting for this type of informative censoring is tantamount to estimating the causal effect of ATP attendance on an outcome if, contrary to the fact, all applicants had remained engaged in research rather than followed their censoring history. We model the exit decision as a function of the same pre-application covariates used to model selection into treatment, and compute weights corresponding to the probability of exit given these observables. The final weight, obtained by multiplying the weights corresponding to the inverse probability of treatment and inverse probability of censoring, is the probability an applicant would have followed his own treatment and censoring history, conditional on observables. We label this methodology Inverse Probability of Treatment and Censoring-Weighted (IPTCW) estimation in what follows.

Selection on unobservables. Despite a long list of observable covariates to predict selection into the ATP, IPTW estimation does little to address the threat to identification due to factors unobservable to the econometrician. The time period of the study suggests an instrumental variable approach based on draft eligibility, as in Angrist (1990). However, medical school graduates, having already deferred their service for educational purposes, were not, in effect, eligible to participate in the lottery (Crowell, 1971; Rousselot, 1971). Table D1 in Appendix D verifies that having one's number called in the lottery does not help predict ATP attendance.

Estimation procedure. Many of the outcomes we study, including publication counts and NIH grants awarded, are skewed and non-negative with a large mass point at zero (see Figures B2a, B2b, and B2c). For example, 426 (13.9%) of the applicants do not publish after their training; approximately two thirds of the sample never receive any NIH grant funding over the career. Following a long-standing tradition in the study of scientific and technical change, for these skewed outcomes we present Poisson quasi-maximum likelihood (hereafter QML) estimates (Santos Silva and Tenreyro, 2006). 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). QML (i.e., "robust") standard errors are computed using the outer product of the gradient vector (and therefore does not rely on the Poisson variance assumption).

4 Results

The exposition of the econometric results proceeds in stages. We first explore empirically the determinants of selection into the ATP. Using the predicted probabilities from these models as regression weights, we then report estimates of the effect of ATP attendance on (i) career choice outcomes; (ii) research productivity (including trainee mentorship outcomes); and (iii) research style outcomes. Finally, we perform a battery of robustness tests to probe the plausibility of the unconfoundedness assumption in our context.

4.1 Selection into the NIH ATP

We model the likelihood of selection in a logit framework using an extensive list of covariates observed at the time of selection (Table 2).[11] We capture the research orientation

[11] In fact, most of these factors might have been observed at the initial selection stage (e.g., medical school attended) while for others the timing is more ambiguous as they might become known to the applicant

of the medical school and intended internship hospital for each applicant with the NIH funding that accrue to principal investigators in these institutions. We also include an indicator variable for applicants who received a PhD before they applied, and an indicator variable for election to the $A\Omega A$ Honor Medical Society. The most informative indicator of research promise is probably demonstrated engagement in research activities, as ascertained by an applicant's list of scientific works published, or soon-to-be-published at the time of application. We weight each of these student publications by the impact factor of the journal in which they appeared as a crude quality adjustment (raw counts produce similar results).

Columns 1a and 1b report logit coefficients and find the signs for most of the covariates are in the expected direction. Relative to applicants without publications and at the mean of all other covariates, computed marginal effects suggest applicants with one publication are 7% more likely to be selected; those with two publications or more, 20% more likely.

Estimates in column 1c correspond to the results of a cross-fit partialing-out lasso logit procedure with ten folds, as described in Chernozhukov et al. (2018). The specification includes all the covariates mentioned above, plus a full suite of medical school indicator variables and a full suite of internship hospitals indicator variables, for a total of 372 covariates, 151 of which the procedure selects for inclusion as control variables. This procedure allows for statistical inference to be performed on five covariates of interest also included in the specification in column 1b, enabling the coefficients and standard errors to be compared across columns.[12]

Columns 2a, 2b, and 2c perform a similar exercise, but the response variable is not selection in this case, but rather exit from research at the end of training. The signs of the coefficient estimates for the predictive covariates are flipped, relative to the specifications in columns 1a, 1b, and 1c.

The specifications used to compute selection probabilities and regression weights for each applicant depart ever so slightly from those in columns 1c (for the selection weights) and 2c (for the informative censoring weights). Since the estimation of the propensity score is solely a prediction exercise, we favor an abundance of explanatory variables in these models. Our least restrictive specification includes 94 fixed effects for medical schools and 238 indicator

between the first and second stage of the ATP selection process (e.g., intended internship hospital, accepted or forthcoming journal publications).

[12] Note that medical school and internship hospital funding variables are not separately identified from the fixed effects and drop out of the specification. The χ^2 test statistic (*i.e.*, the Wald test of the hypothesis that the coefficients of these five covariates are jointly equal to zero) is equal to 78.85 ($p < 0.01$).

variables for intended internship hospitals. We constrain the model to include the same variables as the specification in column 1c and 2c as well as the inverse hyperbolic sine of medical school and internship hospital NIH grant funding. The other variables are selected via a logit procedure with a lasso penalty term, using ten-fold cross-validation to prevent overfitting the data. The predicted probabilities from this model are used to generate the benchmark set of lasso weights used below to estimate treatment effects.[13,14]

Table 1a confirms that pre-application covariates appear balanced across treated and control observations in the sample appropriately weighted using the fitted selection probabilities to construct the selection weights according to the method described in Section 3.3 and Appendix F.

4.2 Career choice

Table 3 reports estimates of the treatment effect of ATP attendance on career outcomes. For each outcome (which differ across rows), the first column reports the naïve cross-sectional estimate. The remaining columns report the average treatment effect (ATE) and the average treatment effect on the treated (ATET) using inverse probability of treatment and censoring lasso weights (computed using the model in Table 2 columns 1c and 2c).

The first two rows of Table 3 report the ATP effect on the length of the training period as well as the length of the career overall. Each estimate in the table corresponds to the coefficient on a treatment indicator variable (and its associated robust standard error) from a Poisson model where the outcome of interest is regressed on an indicator variable for holding a PhD degree at the time of application and a full suite of medical school graduation year effects in addition the treatment variable.

Exponentiated coefficients are presented; subtracting one yields a magnitude interpretable as an elasticity. For example, the estimates in the first cell of Table 3 imply that ATP attendees spend $100 \times (1.091 - 1) = 9.1\%$ longer in training than non-attendees—an additional

[13] We test the quality of our predictions by splitting the sample into a prediction subsample (2,460 or 80% of the observations) and a hold-out sample (615 or 20% of the observations). The out-of-sample deviance ratio (a measure of goodness of fit for logit models) is equal to 0.70 of the corresponding in-sample value, which is acceptable.

[14] As a robustness exercise, we repeat our analysis using logit weights computed using the model in Table 2, columns 1b and 2b (Figure B3, Tables B4a and B4b). Note that the correlation between the predicted selection probabilities from column 1b and that of the model with lasso regularization is 0.919. As a result, the magnitudes and precision of the IPTCW estimates are not very sensitive to the choice of weights.

six months on average. This is a meaningful yet rather small increase relative to the time of commitment of the ATP (two years). It underscores the extent to which our results pertain to the effect of the content of training, rather than to the mere fact that training was received. We also find that NIH training reduces slightly the length of the overall post-independence career, but the effect is small (between 1 and 2%, or seven months on average), and imprecisely estimated in some specifications.

The next six rows of Table 3 pertain to the effect of the program on the choice of career. We report the marginal effects from logistic regressions of these career choice indicators on the treatment indicator and our usual set of controls. Across columns, we observe that attending the ATP greatly increases the likelihood of embarking on an academic or research career. For instance, using the average treatment effect estimated using lasso weights, the marginal effect of starting in academia is 0.11, which corresponds to an odds ratio of 1.77. The program increases the probability of a research-focused initial job even more (the marginal effect is 0.17, which translates into an odds ratio of 2.17) for treated physicians, relative to controls. The effects are also persistent, with similar magnitudes observed when analyzing the program's impact on end-of-career positions. Conversely, attending NIH ATP appears to make it markedly less likely to choose a clinical career (an odds ratio of 0.46).

We also create a composite outcome for joining the biomedical research elite over the course of one's career, which we define as either (i) receiving the Nobel Prize; (ii) being elected to the National Academy of Sciences or the National Academy of Medicine; (iii) being appointed Investigator of the Howard Hughes Medical Institute; or (iv) getting a MERIT designation from the NIH in at least one R01 grant cycle. Only 173 (5.6%) of the applicants belong to this select group by career's end (7.7% of the attendees; 2.2% of the non-attendee controls). Adjusting for selection and censoring based on observable covariates dampens somewhat this difference: the average treatment effect corresponds to an odds ratio of 1.77.

4.3 Research outcomes

Whereas Table 3 focused on the effect of NIH training at the extensive margin (i.e., the choice to begin a research career or to stay in one), Tables 4 and 5 hone in on the effect of the program at the intensive margin (the intensity of research effort over the career, as it is being converted into publications, patents, and grants).

Table 4 reports estimates of the treatment effect of ATP attendance on various metrics of research output over the career. Each outcome variable has been constructed to exclude output that results from research undertaken as a student or a trainee: they correspond to research output for the entire post-training (i.e., "independent") career. We consider nine different outcomes: publication count; publication count excluding those where the applicant is in the middle of the authorship list;[15] cumulative citation count accrued by 2015; USPTO patent count (by 2016); count of references to the scientist's publications appearing on the front page or within the body of patents (Marx and Fuegi, 2020); cumulative NIH grant funding received as a principal investigator; cumulative NIH R01 grant funding received as a principal investigator; count of trainees who go on to receive NIH R01 funding during their own independent careers; and the amount of NIH R01 funding accrued by these trainees.

Synthesizing the results across rows and columns of Table 4, a number of patterns emerge. First, the magnitude of the treatment effects are large, even when they filter out the effect of selection and censoring under the maintained assumption of unconfoundedness. Using the lasso weights, for example, the ATE for publications corresponds to an increase of 63.9%, and the ATET to an increase of 66.8%. Second, modeling selection based on observable covariates does shrink the magnitude of the estimated effects by 25 to 50%, depending on the outcome. Third, the ATE and ATET typically have similar magnitudes, which is logical since control scientists are drawn from the same underlying population. All estimates are precisely estimated, although the ATET specification for patents is only significant at the 10% level.[16]

Citation analysis. The estimates for the effect on overall citations in Table 4 conflate the effect of treatment on the quantity of output with the effect of treatment on the quality of output. Table 5 sheds light on the effect of NIH training on citation impact (a reasonable proxy for publication quality) specifically. For each publication, we use the *Web of Science* to ascertain its percentile in the vintage-specific article-level citation distribution.[17] This makes it possible to meaningfully aggregate, for each applicant, the number of his post-

[15] 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 (Dance, 2012; Sauermann and Haeussler, 2017). Therefore, the first- and last-authored publications correspond to those associated most closely with each applicant.

[16] This is not entirely surprising since applicants in clinical research careers are at very low risk of patenting (only 20% of the physicians on the sample are awarded at least one patent over the course of their career). In contrast, all applicants in the sample are at risk of publishing.

[17] When referring to the vintage-specific, article-level distribution of citations, the relevant universe to compute quantiles is not limited to the articles authored by scientists who belong to our applicant sample.

training publications whose eventual impact falls above the j^{th}-percentile of the citation distribution, even though these publications might have appeared at different times. The structure of Table 5 is otherwise identical to that of Table 3.

The first row of Table 5 replicates the first row of Table 4, with the caveat that we exclude from the publication count variable those for which citations are not available because they appear in a journal indexed by *PubMed* but not the *Web of Science*.[18] The next five rows progressively restrict the count to those whose citations put them above an impact percentile threshold: above the 50^{th}, above the 75^{th}, above the 95^{th}, above the 99^{th}, and above the 99.9^{th} percentile. Looking across rows, the magnitude of the treatment effects increases slightly as one moves up the tail of the impact distribution (except when focusing on the one in a thousand "citation hits"). The more important conclusion is that ATP attendance increases dramatically the number of low-impact as well as the number of high-impact publications over the career.

4.4 Research style

Table 6 examines the impact of NIH training on the style of the research published by applicants to the ATP. Since the style measures cannot be computed absent publications, we limit the analysis in this section to the 2,584 applicants (1,730 attendees and 854 non-attendees) who publish at least once in the post-training phase of the career.[19] The effect on the overall number of publications for the restricted sample of publishers appears in the first row of Table 6 as a benchmark.

A hallmark of the training received at NIH was exposure to laboratory research for young physicians that might have had only limited exposure to the bench as undergraduates or medical school students (and might be unable to receive that style of training in postgraduate fellowships outside of NIH), with an emphasis placed in the oral history on facilitating the "bench to bedside" transition of translational research. Recall that we partition the bibliome into four mutually exclusive styles—basic science, translational medicine, clinical trials, and "other" clinical. The results imply that the program increases output regardless of style,

Rather, the relevant universe includes the entire set of 17,312,059 articles that can be cross-linked between *PubMed* and the *Web of Science*.

[18] These account for 13,853 of 192,785 (7.2%) of all post-independence original research publications for the sample of applicants.

[19] The inverse probability of treatment and censoring weights are recomputed on the restricted sample to take into account the fact that the publication constraint disproportionately drops unsuccessful applicants from the data.

but not evenly. The effect on the number of basic science publications is unambiguously the largest in magnitude, followed by translational and clinical trial publications, with the "other clinical" experiencing only modest and imprecisely estimated increases.[20] We also find that relative to controls, treated physicians publish much more in six high-impact journals prominently advertising a translational focus. In addition, attendees both greatly "inspire" clinical researchers to further develop their translational work, and "stand on translational shoulders" by publishing clinical trials that backward-reference translational articles. Finally, we find that the NIH ATP increases published output that will eventually be cited in one or more USPTO patents.

Considered as a whole, these results points to a durable intellectual imprint associated with the training received at NIH. Some of the trainees became bench scientists, indistinguishable in their output from PhD-holding scientists trained in biology or other basic science departments. Harold Varmus, who went on to win the Nobel Prize in 1989 for his discovery of oncogenes with J. Michael Bishop, is an exemplar of the subset of trainees who leveraged their training to embark on a career at the laboratory bench. Many others, however, did not forsake clinical work completely, but rather acquired in Bethesda an approach to clinical research that was informed by basic research advances, seeding academia with a new generation of who saw themselves as "physician-scientists" rather than "clinician-researchers."

4.5 Mechanisms

It is likely that attending the NIH ATP may impact career and research trajectories through multiple mechanisms, including skill building, signaling, status, peer, and network effects, or instilling values and aspirations (Argote and Fahrenkopf, 2016). Distinguishing between these mechanisms is difficult with the data available, and indeed more than a single mechanism might be responsible for the treatment effects we estimate.

It is notable that many physicians in the control group had exposure to research opportunities outside of the NIH; there was only a small difference in total training time compared to ATP attendees relative to the length of the program. This suggests that the NIH treatment entails more than mere exposure to research. In line with this, we repeat our main analysis, but exclude the 218 applicants who did not attend the ATP and started their independent

[20] Estimating these four specifications jointly enables us to compare the magnitudes explicitly. χ^2 tests strongly reject the hypothesis that the coefficient for basic science is equal to any of the other three categories ($p < 0.01$). Similarly, we can reject the hypothesis that the coefficient for translational medicine and clinical trials are equal to the coefficient for "other clinical" articles. However, we fail to reject the hypothesis that the translational medicine and clinical trial coefficients are in fact equal.

career immediately upon finish residency training, those individuals in the control group least likely to have had substantive research exposure during postgraduate training. The treatment effect magnitudes using this control subsample are, if anything, higher than those observed when using the entire sample (Appendix B, Table B5). This reinforces our contention that the treatment effect should be interpreted as the effect of receiving NIH training relative to "traditional training" and emphasizes the importance of the content rather than the quantity of training received.

Dose-response relationship. Table B6 in Appendix B reports the results of an analysis contrasting the effect of different levels in the intensity of treatment, as proxied by the number of years spent in the ATP. Within the set of 1,929 attendees, 12 (0.6%) spent a year or less at NIH, which we interpret as reflecting the decision to quit the program and receive training elsewhere; 1,321 (68.5%) spent exactly two years as trainees; and 596 (30.9%) three years or more.[21] In these analyses, we model ATP attendance as a multi-valued treatment (Imbens, 2000), and use an ordered logit specification to generate inverse probability of treatment weights. The results uncover a strong dose-response relationship. Across several outcomes, "quitters" and non-attendees exhibit similar outcomes (with the caveat that the effect of quitting is very imprecisely estimated). The effect of spending an additional year within the program is large, and precisely estimated. For example, relative to non-attendees, those staying 3 years publish more over their careers (106% vs. 49%), gather more citations (153% vs. 42%) and are more likely to enter a research job after training (26% vs. 17%) than those staying only the two years necessary to fulfill their service obligation. Once again, we must interpret these results with a great deal of caution, since exposure length is endogenous, and after two years, preceptors are presumably better able to ascertain correctly the research potential of a trainee. While not rejecting selection as a plausible mechanism, this dose-response relationship appears inconsistent with an interpretation of the results based on signaling or status, since it is unlikely that additional years spent in the program would shift future employers' perceptions, or elevate one's status even more in the minds of collaborators, funders, editors, and referees. In addition, the research style evidence seems hard to reconcile with a simple status or signaling story.

Research independence. Another potential mechanism is research independence during training (Shibayama, 2019). This emphasis was reflected in the oral histories of the ATP (see section 2.1). A particular lens on research independence is to examine the extent

[21] This last category includes a small set of about sixty attendees who transitioned from the ATP to another postdoctoral fellowship within NIH, before securing a permanent position.

to which NIH trainees "outperform" their mentors once they leave the nest and become responsible for their own agenda. To provide evidence on this point within the constraints of our data, we first identify the peak vintage-adjusted citation percentile achieved by articles published during training for each trainee. Then, looking only at last-authored publications during career independence, we record the number of articles the former trainee published that exceeded this citation benchmark. Despite having higher peak citation percentiles during training, ATP attendees more frequently exceed their training peaks, relative to non-attendees (Tables B7a and B7b).

Coauthorship-driven peer effects. ATP Fellows could avail themselves to a much broader peer community than in the typical laboratory or clinical fellowship where non-attendees might have completed training, and the oral histories are replete with mentions of the concentration in talent at the NIH brought about by the draft (see Appendix E). ATP attendees may have pushed their colleagues to work harder, to hold higher internal standards of scientific excellence, or helped instill values such as the inherent worth of translational investigations. Unfortunately, while we can observe the cohort of each ATP trainee, we cannot do the same for the non-attendee controls. But we can shed light on the importance of a particular variety of peer effects, those driven by coauthorship.

In Appendix B, Table B8b, we estimate the total number of coauthored papers using a Poisson model with an offset for the total number of publications after career independence. The coefficients are small in magnitude and not statistically significant when comparing all ATP attendees to non-attendees (except for the subset of ATP attendees who stay at NIH after the completion of their training). While we do not find evidence supporting a large role for coauthorship-driven peer effects, this does not preclude an important role for other types of peer effects.

In summary, it is not feasible, within the limitations of our data, to unambiguously pin down a set of mechanisms for the treatment effect magnitudes we estimate. However, the collage of evidence above, together with the results on research style (Section 4.4) imply that skill-based explanations must have played a meaningful role in driving the outcomes we observe. In the conclusion, we argue that in spite of the tentative nature of the evidence regarding mechanisms, some of the ATP's distinctive features provide clues to policy makers as they design training programs adapted to the challenges faced by the 21^{st} century scientific ecosystem.

4.6 Robustness analyses

We perform a number of robustness checks to probe the sensitivity of our estimates to alternative modeling assumptions and subsamples. Recall that in addition to unconfoundedness, the validity of IPTW estimates requires common support. Figure 2 displays the histogram corresponding to the predicted probabilities generated by the selection model in column 1c of Table 2. One can readily observe that the common support assumption is violated in the tails: our model predicts a high probability of selection for very few controls, and low probability of selection for very few treated applicants. The first three columns of Table 7a vary the extent of winsorization for the regression weights: no winsorization (as in Table 4), winsorization at the 5^{th} and 95^{th} percentiles of the distribution of lasso weights; and winsorization at the 10^{th} and 90^{th} percentile of the distribution of lasso weights. The magnitudes of the average treatment effect (corresponding to a single outcome, the number of post-training publications) increases slightly. The violation of the common support assumption is therefore not a first-order concern to assess the robustness of our results.

Rather than weighting by the inverse probability of treatment, the next set of estimates uses coarsened exact matching (Iacus et al., 2011) to match attendees and non-attendees on a handful of covariates: year of medical school graduation, medical school attended, and quintile of the distribution of the pre-application publication count, weighted by journal impact factor. Any treated applicant for whom we cannot find a matched control based on this list of pre-application covariates is simply dropped from the estimation sample. We find that the estimated treatment effect is similar in magnitude to that reported earlier (Table 4).

The last set of two columns in Table 7a focuses on the subset of 1,837 applicants (59.7% of the sample) who had little—if any—research preparation at the time they applied for the program, as ascertained by a lack of any published output. It is of course possible that interviewers were able to divine research potential at the second stage of the selection process, but they would not have had a strong evidentiary record to back up their intuition. The results show that the magnitude of the average treatment effect is just as high, if not higher, in this subpopulation.

Table 7b reports estimates using the "post-double-selection" lasso (hereafter pds-lasso) estimator due to Belloni et al. (2014). This estimator uses the lasso to select covariates to predict both the treatment and the outcome variable, and then estimates the treatment effect of interest by the linear regression of the outcome on the treatment variable and the union of the set of variables selected in the two variable selection steps. The resulting estimator is "doubly robust" in that it allows for imperfect variable selection in either (but not both) of the covariate selection steps. Since the theoretical properties of the pds-lasso estimator have been demonstrated for a linear model, we apply it to our data using ordinary least squares to model the impact of the NIH ATP on the count of post-training publications.[22] The estimates yielded by this procedure are once again large in magnitude, very similar to those associated with IPTW estimation using OLS, and precisely estimated. The point estimate of 26 extra publications, corresponds to 64% of the raw mean difference in the number of publications between attendees and non-attendees.

We also use the bounding technique recently proposed by Oster (2019) to gauge the sensitivity of our results to a failure of the unconfoundedness assumption. The intuition behind this approach is that the stability of the coefficient for the treatment effect when varying the set of control variables included in the model, scaled by movement in R^2, provides information about the potential impact of unobserved covariates. To generate these bounds, the analyst must assume proportionality between the covariances of the outcome with observed and unobserved covariates, and posit a maximum value for R^2 if the regression could include all observed and unobserved covariates. Oster's technique generates a bound δ, the covariance ratio that would be required to reduce the magnitude of the treatment effect to zero. Table 7c reports the results of this exercise for a number of research outcomes. In all cases, δ is far above one, the threshold value recommended by Oster to suggest robustness to the influence of unobservable covariates.

Appendix B includes a number of other robustness checks and ancillary analyses, including: isolating the effect of informative censoring from selection into treatment (Table B10), dynamics of treatment effect over time (Figure B5a and B5b), evidence of imprinting during training (Table B11), heterogeneity in treatment effect by year of program attendance (Table B12), and heterogeneity in treatment effect by program track within the ATP (Table B13).

5 Conclusion

We examine the role of early career exposure to research on sorting into the "ideas sector" of the economy, as well as research trajectory and productivity within this domain. The NIH ATP had a large impact on attendees' careers on both the intensive and extensive

[22] We also use the inverse hyperbolic sine function to transform the publication count. This generates estimates that can approximately be interpreted as elasticities, and therefore compared to those presented in Table 4.

margins. Attendees entered research positions at higher rates after training and remained in them for longer. They not only published more and earned more grant funding, their influence persists through training more second generation scientists and their work was more impactful as measured by citations. More specifically, ATP attendees acquired at NIH a more "translational" style of research, with a greater focus on the bench-to-bedside transition. Remarkably, these changes were sustained throughout their subsequent careers. It is notable that, while there are more "superstars" among ATP attendees than in the set of non-attendee controls, the average physician showed a substantial treatment effect as well. All in all, it is a remarkable impact for a two- to three-year training experience.

Our conclusions depend on the maintained assumption that, conditional on an extensive list of covariates observable at the time of application, selection into the program was essentially random. At first blush, this would appear to be an untenable assumption. While we have adopted a variety of econometric strategies to minimize omitted variable bias, we recognize that at least some of our results could be explained by factors observed by the scientists in charge of selecting the trainees, but not by the econometric analyst. Yet, the institutional setting and the details of the selection process suggest that decision-makers were equally unaware of whom, among the applicants, was decidedly poised for research greatness.

Our control group includes only those who have also applied to the program, which eliminates interest in the program as a potential omitted variable (Jones et al., 2019). In addition, the set of non-attendee controls consists exclusively of those who reached the final interview stage for program admittance and are therefore already highly selected. While we would of course prefer to have interview notes to model the influence of unobservable covariates directly, a large literature suggests that unstructured interviews provide only limited additional information, relative to what is observable on a curriculum vitæ (Dana et al., 2013; Huffcutt et al., 1996; McDaniel et al., 1994; Wiesner and Cronshaw, 1988; Wright et al., 1989). In fact, psychological research has shown that the addition of noisy signals may in fact impair the quality of decision making (Hall et al., 2007; Nisbett et al., 1981). Our reading of this literature leads us to doubt that the unstructured NIH ATP interviews enabled the selection of individuals poised for research greatness. Indeed, medical education is one of a handful of settings where the limited usefulness of interviews has been documented in the field (Milstein et al., 1981).[23] In line with this literature, the oral histories corroborate the difficulty faced by the interviewers in discerning the scientific potential of applicants at

[23] For instance, the University of Texas Medical School at Houston was forced to admit an additional 50 students, all of whom were initially rejected for admission post-interview, due to a legislative decree in 1979;

such an early career stage. Finally, the evidence on research style does not appear to be consistent with the view that selection alone accounts for the results. It strains credulity that the demand side of this labor market might have been able to evaluate aptitude for translational research specifically, in addition to more general research abilities.

Many of the ATP alumni's oral histories evoke the feeling of "being in the right place at the right time." In light of these accounts, the sociological concept of imprinting offers a powerful lens to interpret our results. This stream of research finds that organizations and individuals often exhibit a sensitive period, during which they are susceptible to external influences and come to reflect aspects of this environment, and these aspects can persist despite subsequent environmental changes (Marquis and Tilcsik, 2013; Stinchcombe, 1965). While much of the work on imprinting has focused on firms, there is evidence that imprinting also occurs in the context of individual careers (Baron et al., 1999; Boeker, 1988; Burton and Beckman, 2007; Hannan et al., 1996; Higgins, 2005). During career imprinting, individuals absorb a set of capabilities, connections, and cognitive models from one employer which persist as they change employers later on. Careers are more likely to exhibit the characteristics of an early imprint when their current environment allows them to be surrounded by colleagues with the same imprint, offers them considerable freedom in how they might express an imprint, and if they believe the imprint contributed to prior success (Higgins, 2005). The NIH ATP and the academic medicine context would appear particularly conducive to career imprinting: not only was the ATP an intense experience early in the career, when an imprint is more likely to be absorbed, but the program also had many alumni who seeded the expansion of U.S. Medical Schools in the period immediately following the end of the Vietnam War.[24] Finally, academic research offers a considerable degree of leeway to investigators in structuring the direction and style of their research, and the senior NIH investigators who had acted as mentors to the ATP trainees during the program exemplified the creative use of this autonomy.

these students had no meaningful difference in clinical performance, academic performance and honors, or attrition at either the end of medical school or the first year of postgraduate training (Devaul et al., 1987).

[24] Between 1975 and 2005, the number of faculty members at US Medical Schools increased by a factor of more than two (AAMC Data Book, various editions; Jolly, 1988).

The pool of ATP applicants is not diverse by today's standards.[25] This may pose a challenge to external validity if members of different socioeconomic groups respond differently to the mechanisms of the ATP treatment effect. In particular, interventions that provide a status boost to young scientists can have very different impacts along gender lines (Graddy-Reed et al., 2019). While status is one potential mechanism for the NIH ATP treatment effect, the "dose-response" relationship between the length of training at the ATP and career outcomes argues strongly against an interpretation of the effect mostly, or solely reflecting status considerations. Rather, this evidence is consistent with the idea that trainees are durably imprinted with specific research skills during their stay at NIH. If this view is correct, then there is less reason to fear the findings would not be observed in a more demographically-balanced cohort of trainees.

In light of the unique historical circumstances within which physician research training took place at NIH during the period of our study, we must exercise caution to suggest wider policy implications.[26] Certainly, part of the effectiveness of the ATP in turning physicians into researchers owes much to the extreme concentration of talent in one institution that was facilitated by the Vietnam War. The effects of the ATP may have been large and long-lasting precisely because the exposure received was intense. Yet, this program provides an existence proof for the proposition that it is possible to design interventions to turn individuals who in the main would not have had scientific careers into frontier researchers. This stands in contrast with many other active labor market policies often studied by economists. The effects of these programs are typically modest in magnitude, and their effects relatively transitory (Heckman et al., 1999). Conversely, the labor market effects of military service appear to mostly correspond to loss of experience, as the earnings profiles of veterans and non-veterans converge relatively quickly (Angrist, 1990; Angrist et al., 2011).

There have been attempts to recreate the "hot house" environment that characterized the intramural campus of the NIH in the 1960s and 1970s (Rubin, 2006). But which characteristics of the NIH ATP were instrumental in its ability to push attendees towards the heights of the biomedical research elite? The unique set of circumstances is unlikely to occur again, and it would be depressing to suggest that the exigencies of wartime are a necessary

[25] We came across two African-American physicians in the entire sample. Similarly, we found 36 female applicants (include 2 foreign medical graduates and 1 lost to follow up) in the NIH index cards (15 attendees, 21 non-attendees), who may have been discriminated against in the application process because any spot occupied by a woman entailed that a male physician would serve in the armed forces, possibly in the Vietnam theater (although in fact few among our control appear to have served in South Asia, if they did serve at all).

[26] Appendix E contains a discussion of the estimated program costs and return on investment.

condition for the design of effective scientific training programs. Such pessimism is not warranted. We emphasize three features of the ATP relevant for the design of training programs today, within and beyond the setting of biomedicine. First is the timing of training receipt, which for many ATP attendees was their first serious engagement with scientific investigation. In many respects, the ATP was more akin to a "pre-doc" than a graduate school or postdoctoral experience. Second is the size of each cohort. The ATP cohorts were much larger than those in the typical Medical Scientist Training Program or other research fellowships. This may intersect in important ways with the role of peer effects, facilitated by the concentration of talent in one location during the NIH ATP. Third, the ATP stressed building independence. This is very different from the modern setting, where typically all papers are automatically coauthored with the principal investigator, there is often little scope to deviate from the principal investigator's research agenda, and many budding scientists linger in training, typically in a sequence of post-doctoral positions (Kahn and Ginther, 2017).

Yet, despite these distinctive features, it is difficult to offer firm guidance for scientific training programs. Our evidence unfortunately does not allow us to empirically isolate the individual mechanisms explaining the effect of ATP attendance. This is an opportunity. The very success of the ATP suggests policy makers should experiment with design features that were its hallmarks. We conclude with a call for more systematic and rigorous evaluation of training programs. It is an unfortunate paradox that Randomized Control Trials (RCTs) are a staple of the biomedical research enterprise, and yet seem to be viewed as out of place in the context of funding and training policies. In our view, the lowest hanging fruit available to designers of training programs—especially those with more applicants than available seats—is to build in evaluation in the design phase, instead of treating it as an afterthought.

References

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

Association of American Medical Colleges. 2016. "Diversity in Medical Education: Fats & Figures 2016." Washington, D.C.: AAMC.

Anfinsen, Christian B. 1963. "History of the Research Associate Program." Office of NIH History.

Angrist, Joshua D. 1990. "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records." *The American Economic Review* **80**(3): 313-336.

Angrist, Joshua D., Stacey H. Chen, and Jae Song. 2011. "Long-term Consequences of Vietnam-Era Conscription: New Estimates Using Social Security Data." *American Economic Review: AEA Conference & Proceedings* **101**(3): 334-338.

Argote, Linda, and Erin Fahrenkopf. 2016. "Knowledge Transfer in Organizations: The Roles of Members, Tasks, Tools and Networks." *Organizational Behavior and Human Decision Processes* **136**: 146-159.

Austin, Peter C., and Elizabeth Stuart. 2015. "Moving Towards Best Practice When Using Inverse Probability of Treatment Weighting (IPTW) Using the Propensity Score to Estimate Causal Treatment Effects in Observational Studies." *Statistics in Medicine* **34**(28): 3661-3679.

Azoulay, Pierre, Waverly Ding, and Toby Stuart. 2009. "The Effect of Academic Patenting on the Rate, Quality, and Direction of (Public) Research Output." *Journal of Industrial Economics* **57**(4):637-676.

Baron, James N., M. Diane Burton, and Michael T. Hannan. 1999. "Engineering Bureaucracy: The Genesis of Formal Policies, Positions, and Structures in High-Technology Firms." *Journal of Law, Economics, and Organization* **15**(1): 1-41.

Baskir, Lawrence N., and William A. Strauss. 1978. *Chance and Circumstance: The Draft, The War, and the Vietnam Generation.* New York, NY: Vintage Books.

Bell, Alexander M., Raj Chetty, Xavier Jaravel, Neviana Petkova, and John Van Reenen. 2019. "Who Becomes an Inventor in America? The Importance of Exposure to Innovation." *Quarterly Journal of Economics* **134**(2):647-713.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014. "Inference on Treatment Effects after Selection among High-Dimensional Controls." *Review of Economic Studies* **81**(2): 608-650.

Berry, Frank B. 1976. "The Story of 'The Berry Plan'." *Bulletin of the New York Academy of Medicine* **52**(3): 278-282.

Blau, David M., and Bruce A. Weinberg. 2017. "Why the US Science and Engineering Workforce is Aging Rapidly." *Proceedings of the National Academy of Sciences* **114**(15): 3879-3884.

Blume-Kohout, Margaret E., and Dadhi Adhikari. 2016. "Training the scientific workforce: Does funding mechanism matter?" *Research Policy* **45**(6):1291-1303.

Boeker, Warren. 1988. "Organizational Origins: Entrepreneurial and Environmental Imprinting at Time of Founding." In Glenn R. Carroll (Ed.), *Ecological Models of Organizations*, pp. 33-51. Cambridge, MA: Ballinger.

Broström, Anders. 2019. "Academic Breeding Grounds: Home Department Conditions and Early Career Performance of Academic Researchers." *Research Policy* **48**(7):1647-1665.

Burton, M. Diane, and Christine M. Beckman. 2007. "Leaving a Legacy: Position Imprints and Successor Turnover in Young Firms." *American Sociological Review* **72**(2): 239-266.

Butler, Declan. 2008. "Translational Research: Crossing the Valley of Death." *Nature* **453**(7197): 840-842.

Bynum, William. 2012. "What Makes a Great Lab?" *Nature* **490**(7418): 31-32.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. "Double/debiased Machine Learning for Treatment and Structural Parameters." *Econometrics Journal* **21**(1): C1-C68.

Crowell, James A. 1971. "The Procurement of Physicians for the United States Armed Forces." Army War College Working Paper AD-769 594.

Dana, Jason, Robyn Dawes, and Nathanial Peterson. 2013. "Belief in the Unstructured Interview: The Persistence of an Illusion." *Judgment and Decision Making* **8**(5): 512-520.

Dance, Amber. 2012. "Who's on First?" *Nature* **489**(7417): 591-593.

Deming, David J., and Kadeem L. Noray. 2018. "STEM Careers and Technological Change." NBER Working Paper #25065.

Devaul, Richard, Faith Jevey, James Chappell, Patricia Caver, Barbara Short, and Stephen O'Keefe. 1987. "Medical School Performance of Initially Rejected Students." *Journal of the American Medical Association* **257**(1): 47-51.

Fauci, Anthony S. 1989. "In Their Own Words AIDS Oral History Series." edited by Victoria Harden. Office of NIH History, National Institutes of Health.

Fauci, Anthony S. 1998. "Clinical Associates Program Oral History Series." edited by Melissa Klein. Office of NIH History, National Institutes of Health.

Fredrickson, Donald S. 1998. "Clinical Associates Program Oral History Series." edited by Melissa Klein. Office of NIH History, National Institutes of Health.

Freeman, Richard B. 1975. "Supply and Salary Adjustments to the Changing Science Manpower Market: Physics, 1948-1973." *American Economic Review* **65**(1): 27-39.

Ginther, Donna K., Janet M. Currie, Francine D. Blau, and Rachel T.A. Croson. 2020. "Can Mentoring Help Female Assistant Professors? An Evaluation by Randomized Trial." *American Economic Association Papers & Proceedings* **110**(5): 205-209.

Ginther, Donna K., and Misty L. Heggeness. 2020. "Administrative Discretion in Scientific Funding: Evidence from a Prestigious Postdoctoral Training Program." *Research Policy* **49**(4): 103953.

Goldstein, Joseph L., and Michael S. Brown. 1997. "The Clinical Investigator: Bewitched, Bothered, and Bewildered—But Still Beloved." *Journal of Clinical Investigation* **99**(12): 2803-2812.

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

Graddy-Reed, Alexandra, Lauren Lanahan, and Jonathan Eyer. 2019. "Gender discrepancies in publication productivity of high-performing life science graduate students." *Research Policy* **48**(9):103838.

Hall, Crystal C., Lynn Ariss, and Alexander Todorov. 2007. “The Illusion of Knowledge: When More Information Reduces Accuracy and Increases Confidence.” *Organizational Behavior and Human Decision Processes* **103**(2): 277-290.

Hannan, Michael T., M. Diane Burton, and James N. Baron. 1996. “Inertia and Change in the Early Years: Employment Relations in Young, High Technology Firms.” *Industrial and Corporate Change* **5**(2): 503-536.

Heckman, James J., Robert J. Lalonde, and Jeffrey A. Smith. 1999. “The Economics and Econometrics of Active Labor Market Programs.” In Orley C. Ashenfelter, and David Card (Eds.), Chapter 31 in *Handbook of Labor Economics* **3A**: 1865-2097. Amsterdam: Elsevier North-Holland.

Higgins, Monica C. 2005. *Career Imprints: Creating Leaders Across an Industry.* San Francisco: Jossey-Bass/Wiley.

Hill, Ryan Reed. 2018. “Searching for Superstars: Research Risk and Talent Discovery in Astronomy.” Working Paper, MIT.

Hirano, Keisuke, and Guido W. Imbens. 2001. “Estimation of Causal Effects using Propensity Score Weighting: An Application to Data on Right Heart Catheterization.” *Health Services & Outcomes Research Methodology* **2**: 259-278.

Huffcutt, Allen I., Philip L. Roth, and Michael A. McDaniel. 1996. “A meta-analytic investigation of cognitive ability in employment interview evaluations: Moderating characteristics and implications for incremental validity.” *Journal of Applied Psychology* **81**(5): 459-473.

Iacus, Stefano M., Gary King, and Giuseppe Porro. 2011. “Multivariate Matching Methods That are Monotonic Imbalance Bounding.” *Journal of the American Statistical Association* **106**(493): 345-361.

Imbens, Guido W. 2000. “The Role of the Propensity Score in Estimating Dose-response Functions.” *Biometrika* **87**(3): 706-710.

Jacob, Brian A., and Lars Lefgren. 2011. “The impact of NIH postdoctoral training grants on scientific productivity.” *Research Policy* **40**(6):864-874.

Jolly, Paul. 1988. “Medical Education in the United States, 1960-1987.” *Health Affairs* **7**(Suppl 2): 144-157.

Jones, Charles I. 1995. “R&D-based Models of Economic Growth.” *Journal of Political Economy* **103**(4): 759-784.

Jones, Damon, David Molitor, and Julian Reif. 2019. “What do Workplace Wellness Programs Do? Evidence from the Illinois Workplace Wellness Study.” *The Quarterly Journal of Economics* **143**(4):1747-1791.

Kahn, Shulamit and Donna K. Ginther. 2017. “The Impact of Postdocs on Early Careers in Biomedicine.” *Nature Biotechnology* **35**(1): 90-94.

Keiser, Harry. 1998. “Clinical Associates Program Oral History Series.” edited by Melissa Klein. Office of NIH History, National Institutes of Health.

Kerr, William R. 2018. *The Gift of Global Talent: How Migration Shapes Business, Economy and Society.* Stanford University Press.

Khot, Sandeep, Buhm Soon Park, and Jr. W.T. Longstreth. 2011. “The Vietnam War and Medical Research: Untold Legacy of the U.S. Doctor Draft and the NIH “Yellow Berets”.” *Academic Medicine* **86**(4): 502-508.

Kimball, Harry R. 1997. “Clinical Associates Program Oral History Series.” edited by Melissa Klein. Office of NIH History, National Institutes of Health.

Klein, Melissa K. 1998. “The Legacy of the ‘Yellow Berets’: The Vietnam War, the Doctor Draft, and the NIH Associate Training Program.” Manuscript, NIH History Office, National Institutes of Health.

Marquis, Christopher, and Andras Tilcsik. 2013. “Imprinting: Toward a Multilevel Theory.” *The Academy of Management Annals* **7**(1): 193-243.

Marx, Matt and Aaron Fuegi. 2020. “Reliance on Science: Worldwide Front-Page Patent Citations to Scientific Articles.” *Strategic Management Journal* **41**(9):1572-1594.

McDaniel, Michael A., Deborah L. Whetzel, Frank L. Schmidt, and Steven D. Maurer. 1994. “The Validity of Employment Interviews: A Comprehensive Review and Meta-Analysis.” *Journal of Applied Psychology* **79**(4): 599-616.

Milojevic, Staša, Filippo Radicchi, and John P. Walsh. 2018. “Changing Demographics of Scientific Careers: The Rise of the Temporary Workforce.” *Proceedings of the National Academy of Sciences* **115**(50): 12616-12623.

Milstein, Robert M., Leland Wilkinson, Gerard N. Burrow, and William Kessen. 1981. “Admission Decisions and Performance During Medical School.” *Journal of Medical Education* **56**(2): 77-82.

Nathan, David G. 2005. “The Several Cs of Translational Clinical Research.” *Journal of Clinical Investigation* **115**(4): 795-797.

NIH. 1968. “Associate Training Program in the Medical and Biological Sciences at the National Institutes of Health.” Department of Health, Education, and Welfare.

NIH Office of Research Information. 1963. “New Class of 101 New Physicians Join NIH Research Training Programs.” NIH Office of Research Information, National Institutes of Health.

Nisbett, Richard E., Henry Zukier, and Ronald E. Lemley. 1981. “The Dilution Effect: Nondiagnositc Information Weakens the Implications of Diagnostic Information.” *Cognitive Psychology* **13**(2): 248-277.

Oster, Emily. 2019. “Unobservable Selection and Coefficient Stability: Theory and Evidence.” *Journal of Business & Economic Statistics* **37**(2): 187-204.

Park, Buhm Soon. 2003. “The Development of the Intramural Research Program at the National Institutes of Health after World War II.” *Perspectives in Biology and Medicine* **46**(3): 383-402.

Robins, James M., Miguel A. Hernan, and Babette Brumback. 2000. “Marginal Structural Models and Causal Inference in Epidemiology.” *Epidemiology* **11**(5):550-560.

Rhodes, Richard. 1986. *The Making of the Atomic Bomb*. New York: Touchstone.

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

Rousselot, Louis M. 1971. “Doctor Draft.” *Archives of Surgery* **102**(1): 88-89.

Rubin, Gerald M. 2006. “Janelia Farm: An Experiment in Scientific Culture.” *Cell* **125**(2): 209-212.

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

Sauermann, Henry, and Carolin Haeussler. 2017. "Authorship and contribution disclosures." *Science Advances* **3**(11): e1700404.

Shibayama, Sotaro. 2019. "Sustainable development of science and scientists: Academic training in life science labs." *Research Policy* **48**(3):676-692.

Snyder, Thomas D. (Ed.). 1993. "120 Years of American Education: A Statistical Portrait." Washington, D.C.: U.S. Department of Education.

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

Stinchcombe, Arthur L. 1965. "Social Structure and Organizations." In James G. March (Ed.), *Handbook of Organizations*, (pp. 142-193) Chicago, IL: Rand McNally.

Svorencík, Andrej. 2014. "MIT's Rise to Prominence: Outline of a Collective Biography." *History of Political Economy* **46**(suppl 1): 109-133.

Teitelbaum, Michael S. 2014. *Falling Behind? Boom, Bust and the Global Race for Scientific Talent*. Princeton, NJ: Princeton University Press.

Urschel, John, and Louisa Thomas. 2019. *Mind and Matter: A Life in Math and Football*. New York: Penguin Group.

Varmus, Harold. 2009. *The Art and Politics of Science*: W. W. Norton & Company.

Wiesner, Willi H., and Steven F. Cronshaw. 1988. "A Meta-analytic Investigation of the Impact of Interview Format and Degree of Structure on the Validity of the Employment Interview." *Journal of Occupational Psychology* **61**(4): 275-290.

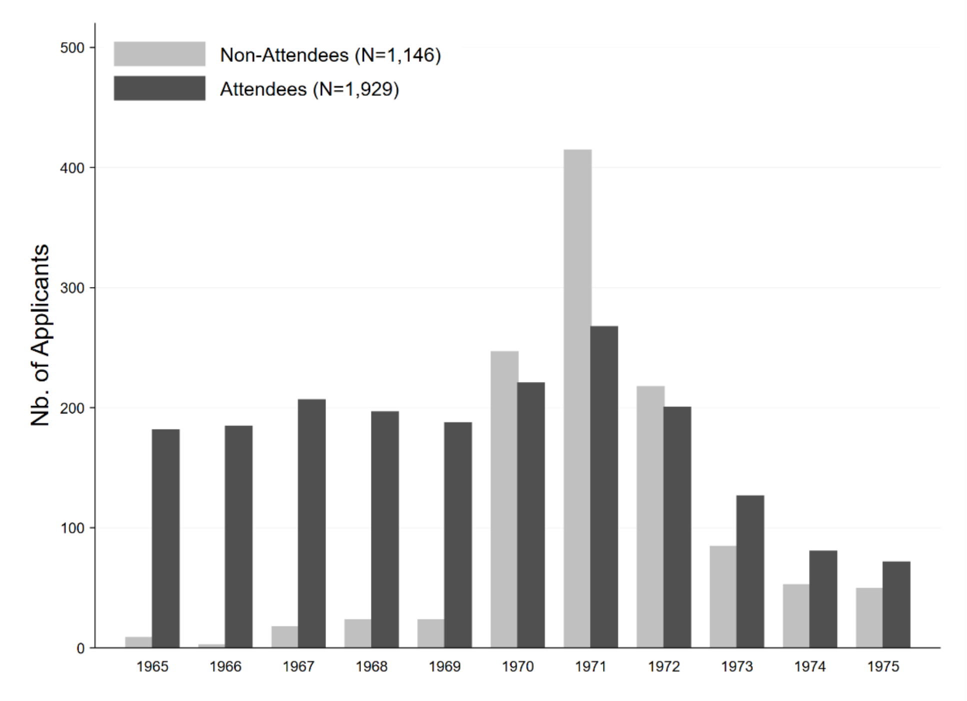
Woolf, Steven H. 2008. "The Meaning of Translational Research and Why it Matters." *Journal of the American Medical Association* **299**(2): 211-213.

Wright, Patrick M., Philip A. Lichtenfels, and Elliot D. Pursell. 1989. "The Structured Interview: Additional Studies and a Meta-analysis." *Journal of Occupational Psychology* **62**(3): 191-199.

Wyngaarden, James B. 1979. "The Clinical Investigator as an Endangered Species." *New England Journal of Medicine* **301**(23): 1254-1259.

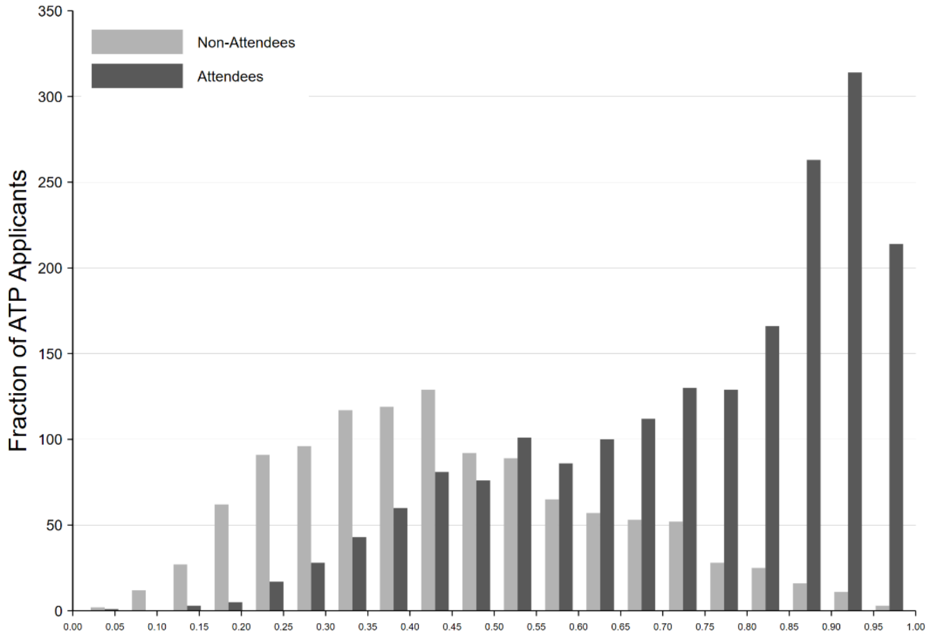
Xu, Stanley, Colleen Ross, Marsha A. Raebel, Susan Shetterly, Christopher Blanchette, and David Smith. 2010. "Use of Stabilized Inverse Propensity Scores as Weights to Directly Estimate Relative Risk and Its Confidence Intervals." *Value in Health* **13**(2): 273-277.

Figure 1. NIH ATP interviewed candidates by year



Note: Number of second-round applicants, by year and treatment status. N=3,075 applicants (1,929 attendees; 1,146 non-attendees). *Sources*: ATP Index Cards.

Figure 2. Predicted probability of selection



Note: Predicted probabilities from the lasso penalized logit procedure described in the last paragraph of section 4.1 of the manuscript.

Table 1a. Descriptive statistics: Pre-application data

	Unweighted sample			Lasso IPTW Reweighting		
	Non-attendees	**Attendees**	***t*-stat**	**Non-attendees**	**Attendees**	***t*-stat**
PhD	0.013	0.036	3.738	0.030	0.027	0.216
	(0.003)	(0.004)		(0.011)	(0.003)	
Age in the Year of Last Application	25.931	26.016	1.596	26.021	26.145	0.923
	(0.042)	(0.033)		(0.056)	(0.139)	
Applies more than once	0.027	0.028	0.154	0.043	0.028	1.434
	(0.005)	(0.004)		(0.010)	(0.004)	
Number of Applications	1.028	1.029	0.093	1.044	1.029	1.426
	(0.005)	(0.004)		(0.010)	(0.004)	
Number of Institutes Applied For	2.948	3.933	11.789	3.501	3.596	0.730
	(0.061)	(0.053)		(0.115)	(0.057)	
Number of Associate Tracks Applied For	1.828	2.068	7.811	1.962	1.991	0.729
	(0.025)	(0.019)		(0.034)	(0.021)	
AΩA Honor Medical Society	0.257	0.383	7.162	0.305	0.328	0.894
	(0.013)	(0.011)		(0.019)	(0.016)	
Pre-ATP Nb. of Publications	0.582	1.005	7.367	0.841	0.846	0.063
	(0.037)	(0.039)		(0.080)	(0.045)	
Pre-ATP JIF-weighted Nb. of Publications	3.288	6.595	6.835	6.107	5.289	0.854
	(0.292)	(0.330)		(0.967)	(0.330)	
NIH Grants for Applicant's Medical School	170.323	207.006	6.879	185.481	189.629	0.428
	(3.792)	(3.438)		(6.033)	(7.441)	
NIH Grants for Applicant's Internship Hospital	90.915	97.153	1.978	90.758	92.300	0.294
	(2.590)	(1.882)		(3.505)	(3.863)	
Attended Harvard Medical School	0.075	0.142	5.614	0.101	0.121	1.238
	(0.008)	(0.008)		(0.014)	(0.008)	
Attended Johns Hopkins School of Medicine	0.047	0.059	1.356	0.043	0.056	1.576
	(0.006)	(0.005)		(0.007)	(0.006)	
Attended Columbia University	0.050	0.044	0.725	0.051	0.042	0.965
	(0.006)	(0.005)		(0.008)	(0.005)	

Note: N=3,075 applicants (1,929 attendees; 1,146 non-attendees). Means, standard errors, and t-statistics are reported; reweighting is performed using average treatment effect inverse probability of treatment weights. T-statistics are calculated using IPTW-weighted OLS regression of the variable of interest on an indicator variable for ATP attendance. Harvard, Johns Hopkins, and Columbia are the three most common medical schools attended in the sample. For NIH grants, original amounts were deflated using the Biomedical R&D Producer Price Index (2015 dollars) and presented in units of millions of dollars. JIF—journal impact factor. *Sources*: ATP Index Cards, PubMed, CGAF.

Table 1b. Descriptive statistics: Career choice

	Non-Attendees		Attendees	
	Mean	Std. Dev.	Mean	Std. Dev.
Deceased	0.075	0.264	0.100	0.299
Years of Post-graduate Training	5.864	1.688	6.425	1.556
Nb. of Career Years (censored in 2017)	37.651	5.805	38.149	6.389
First Job in Academia	0.572	0.495	0.757	0.429
Ends Career in Academia	0.381	0.486	0.546	0.498
Researcher First Job	0.460	0.499	0.694	0.461
Ends Career as Researcher	0.300	0.459	0.519	0.500
First Job in Clinical Practice	0.535	0.499	0.296	0.457
Ends Career in Clinical Practice	0.657	0.475	0.441	0.497

Note: Academia includes both universities/medical schools and research settings such as the NIH or private non-profit institutes (e.g., The Salk Research Institute). Researcher jobs is different from academia in that it includes for-profit industry research positions but excludes clinical university faculty. Clinical practice includes both those in community practice as well as medical school clinical faculty. All variables except years post-graduate training and number of career years are indicator variables. *Sources*: ATP Index Cards, AMA Physician Masterfile, doximity.com, state licensure records, NIH telephone directories.

Table 1c. Descriptive statistics: Research outcomes

	Non-Attendees		Attendees	
	Mean	Std. Dev.	Mean	Std. Dev.
Nb. of Pubs, Training Period	2.400	4.079	6.050	6.389
Career Nb. of Pubs	37.313	80.078	77.773	109.584
Career Citations	1,988	5,345	5,131	10,391
Nb. of Patents	0.657	3.729	1.738	6.569
Career Citations to Patents in Patents	7.506	53.651	20.227	106.080
Career Citations to Pubs in Patents	80.095	347.028	252.029	914.263
NIH Grant Recipient	0.206	0.405	0.442	0.497
Career NIH Grants ($ 2015)	4,511,372	35,192,232	12,436,209	42,898,984
Career NIH R01 Grants ($ 2015)	1,193,642	5,035,673	3,149,951	8,197,320
Nb. NIH-R01-funded Trainees	0.214	0.885	0.758	1.914
Trainee Career NIH R01 Grants ($2015)	1,167,519	6,091,129	4,722,876	14,203,229

Note: Except in the first row, all outcomes should be understood to be restricted to output in the post-training (i.e., independent) phase of the career. NIH grant recipient is an indicator variable equal to 1 if an individual ever received an NIH grant. *Sources*: ATP Index Cards, PubMed, CGAF, USPTO, Marx and Fuegi (2020) “reliance on science” publication-to-patent linkages.

Table 1d. Notable achievements

	Nobel Prize	Natl. Academies Member	Howard Hughes Med. Investigator	NIH MERIT [R37] Awardee
Non-Attendees	0 (0.00%)	14 (1.12%)	0 (0.00%)	14 (1.22%)
Attendees	7 (0.36%)	90 (4.67%)	32 (1.66%)	79 (4.10%)
Total	7 (0.23%)	104 (3.34%)	32 (1.04%)	93 (3.02%)

Sources: ATP Index Cards, CGAF, Nobel Prize, HHMI, and NAS web sites.

Table 1e. Descriptive statistics: Research style

	Non-Attendees		Attendees	
	Mean	Std. Dev.	Mean	Std. Dev.
Basic Science Articles	0.107	0.200	0.199	0.248
Translational Medicine Articles	0.209	0.234	0.273	0.232
Clinical Trial Articles	0.097	0.161	0.107	0.162
Other Clinical Articles	0.467	0.324	0.338	0.292
Articles Appearing in “Translational” Journals	0.012	0.065	0.016	0.039
Inspires Translational Research	0.088	0.135	0.118	0.137
Builds on Translational Research	0.068	0.131	0.078	0.130
Articles Cited in Patents	0.109	0.149	0.162	0.162

Note: N=2,584 scientists (491 scientists with zero publications cited at least once in the independent phase of the career are excluded). Statistics correspond to the fraction of each scientist's work with the corresponding characteristic. *Sources*: ATP Index Cards, PubMed.

Table 2. Modeling selection into the NIH ATP

	Program Selection			***Informative Censoring***		
	Parsimonious Model [Logit]		**Saturated Model [Lasso]**	**Parsimonious Model [Logit]**		**Saturated Model [Lasso]**
	(1a)	(1b)	(1c)	(2a)	(2b)	(2c)
Log(Pre-ATP Nb. of Publications)		0.307**	0.329**		-0.192**	-0.210**
		(0.071)	(0.071)		(0.064)	(0.066)
Ln(NIH Grants for Applicant's Medical School)	0.357**	0.317**		-0.193**	-0.158*	
	(0.090)	(0.091)		(0.067)	(0.066)	
Ln(NIH Grants for Applicant's Internship Hospital)	0.019*	0.017†		-0.031**	-0.029**	
	(0.009)	(0.010)		(0.008)	(0.009)	
PhD	0.932**	0.577†	0.794*	-1.347**	-1.036**	-1.141**
	(0.334)	(0.342)	(0.311)	(0.354)	(0.358)	(0.359)
No Internship	1.962*	1.763*	1.117	-2.716*	-2.583*	-3.936**
	(0.839)	(0.849)	(0.981)	(1.056)	(1.068)	(0.879)
Applies more than once	-0.039	-0.091	0.071	0.076	0.115	-0.028
	(0.300)	(0.295)	(0.273)	(0.246)	(0.249)	(0.242)
AΩA Honor Medical Society	0.688**	0.701**	0.661**	-0.346**	-0.346**	-0.347**
	(0.105)	(0.106)	(0.101)	(0.087)	(0.088)	(0.087)
Constant	-3.278†	-2.649		3.049*	2.329†	
	(1.748)	(1.775)		(1.311)	(1.310)	
Medical School Fixed Effects	No	No	Yes	No	No	Yes
Internship Hospitals Fixed Effects	No	No	Yes	No	No	Yes
Nb. of Non-zero Predictors			151			168
Nb. of Potential Predictors			372			372
χ^2 Test Statistic			67.96			51.23
Pseudo-R^2	0.251	0.265		0.056	0.073	
Log-likelihood	-1,521	-1,493		-1,945	-1,910	
Nb. of Applicants	3,075	3,075	3,075	3,075	3,075	3,073

Note: The dependent variable is an indicator variable equal to one for attendees, zero for non-attendees (first three columns) or an indicator variable equal to one for attendees who exit research immediately after training (last three columns). Estimates are displayed as coefficients from logit specifications. All models incorporate a full suite of medical school graduation year effects; a set of indicator variables for the applicant's age at the time of application; indicator variables for the number of distinct NIH component institutes that received the application; indicator variables for the number of tracks applied to within the Associate Training Program; indicator variables for the number of years between the application and the medical school graduation year; and a series of indicator variables capturing if the applicant (1) intended to postpone his internship until after training, (2) intends to perform his internship abroad, (3) intends to intern in a hospital affiliated with the Veterans Affairs Administration, or (4) has missing information regarding his intended internship hospital. All models except (1a) and (2a) also include an indicator variable for applicants without any publication before application. Estimates in columns [1c] and [2c] correspond to the results of a cross-fit partialing-out lasso logit procedure with ten folds, as described in Chernozhukov et al. (2018). The specification includes all the covariates mentioned above, plus a full suite of medical school indicator variables and a full suite of internship hospitals indicator variables, but only a subset of this list is selected for inclusion (151 out of 372 in model [1c]; 168 out of 372 in model [2c]. In both models [1c] and [2c], a Wald test rejects the hypothesis that the "coefficients of interest" (i.e., those that are constrained to appear in the model, and for which inference is performed) are jointly equal to zero. Robust errors in parentheses ($^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$). *Sources*: ATP Index Cards, PubMed, CGAF.

Table 3. Career choice outcomes

	X-Sect.	Lasso Weights	
	Naive	ATE	ATET
Poisson Estimates			
Years of Post-graduate Training	1.091** (0.012)	1.082** (0.015)	1.069** (0.019)
Nb. of Career Years	0.988† (0.006)	0.988† (0.007)	0.989 (0.008)
Logit Estimates			
First Job in Academia	0.160** (0.018)	0.113** (0.021)	0.085** (0.025)
Ends Career in Academia	0.146** (0.020)	0.132** (0.032)	0.111** (0.027)
Researcher First Job	0.212** (0.018)	0.170** (0.022)	0.153** (0.026)
Ends Career as Researcher	0.216** (0.019)	0.192** (0.030)	0.175** (0.025)
First Job in Clinical Practice	-0.212** (0.018)	-0.171** (0.022)	-0.155** (0.026)
Ends Career in Clinical Practice	-0.215** (0.019)	-0.189** (0.029)	-0.174** (0.025)
Joins the Research Elite	0.056** (0.013)	0.025* (0.012)	0.032* (0.013)
Number of Applicants	3,075	3,075	3,075

Note: Each cell contains an estimate for the treatment effect in a separate regression. The dependent variables are listed in the left-most column. All models incorporate a full suite of medical school graduation year effects as well as an indicator variable for holding a PhD degree at the time of application. Column 2 and 3 perform inverse probability of treatment weighted estimation for first career position and training length outcomes (rows 1, 2, 3, 5, and 7) and inverse probability of treatment and censoring weighted estimation for all other outcomes; the corresponding estimates can be interpreted as the ATE/ATET of NIH training, under the assumption of unconfoundedness. On the first two rows, the estimates stem from Poisson regressions. Exponentiated coefficients are presented; subtracting 1 yields magnitudes interpretable as elasticities. For example, the estimate in the top cell of the first column imply that attendees stay 100×(1.091]-1)=9.1% longer in training, relative to non-attendees; the effect is highly statistically significant. On the next six rows, the estimates stem from logistic regressions. The marginal effects for the treatment indicator are reported. For instance, the coefficient in the third row of the first column implies that attendees are 16.0% more likely than non-attendees to be initially placed in academia after completing their training. Robust errors in parentheses ($^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$). *Sources*: ATP Index Cards, AMA Physician Masterfile, doximity.com, state licensure records, NIH telephone directories.

Table 4. Research outcomes

	X-Sect.	Lasso Weights	
	Naïve	ATE	ATET
Career Nb. of Pubs	1.922**	1.639**	1.668**
	(0.146)	(0.128)	(0.152)
Career Nb. of Pubs, First/Last Authorship Position	1.970**	1.673**	1.701**
	(0.146)	(0.131)	(0.157)
Career Citations	2.316**	1.775**	1.884**
	(0.227)	(0.190)	(0.220)
Nb. of Patents	2.515**	1.635*	1.559†
	(0.521)	(0.368)	(0.406)
Career Citations to Pubs in Patents	3.015**	1.892**	1.938**
	(0.511)	(0.386)	(0.454)
Career NIH Grants	2.558**	1.846*	1.807*
	(0.591)	(0.489)	(0.517)
Career NIH R01 Grants	2.285**	1.668**	1.774**
	(0.352)	(0.296)	(0.340)
Nb. NIH-R01-Funded Trainees	2.410**	1.621*	1.689*
	(0.365)	(0.349)	(0.429)
Trainee Career NIH R01 Grants	2.677**	1.890**	2.039**
	(0.501)	(0.423)	(0.541)
Number of Applicants	3,075	3,075	3,075

Note: Each cell contains an estimate for the treatment effect in a separate regression. All estimates stem from Poisson regressions. The dependent variables are listed in the left-most column. All models incorporate a full suite of medical school graduation year effects as well as an indicator variable for holding a PhD degree at the time of application. Exponentiated coefficients are presented; subtracting 1 yields magnitudes interpretable as elasticities. For example, the estimate in the first cell imply that attendees publish 100×(1.922-1)=92.2% more original articles during the independent phase of their career, relative to non-attendees; the effect is highly statistically significant. Columns 2 and 3 perform inverse probability of treatment and censoring weighted; the corresponding estimates can be interpreted as the ATE/ATET of NIH training, under the assumption of unconfoundedness. Robust errors in parentheses ($^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$). *Sources*: ATP Index Cards, PubMed, Web of Science, CGAF, USPTO, Marx and Fuegi (2020) "reliance on science" publication-to-patent linkages.

Table 5. Publication outcomes, by citation quantiles

	X-Sect.	Lasso Weights	
	Naïve	ATE	ATET
Career Nb. of Pubs, Total	1.951^{**}	1.637^{**}	1.680^{**}
(with citation data available)	(0.151)	(0.130)	(0.156)
Career Nb. of Pubs	2.065^{**}	1.696^{**}	1.735^{**}
Top 50% of the Citation Distribution	(0.169)	(0.146)	(0.175)
Career Nb. of Pubs	2.157^{**}	1.717^{**}	1.771^{**}
Top 25% of the Citation Distribution	(0.189)	(0.162)	(0.194)
Career Nb. of Pubs	2.348^{**}	1.800^{**}	1.884^{**}
Top 5% of the Citation Distribution	(0.247)	(0.202)	(0.240)
Career Nb. of Pubs	2.653^{**}	1.963^{**}	2.176^{**}
Top 1% of the Citation Distribution	(0.347)	(0.284)	(0.322)
Career Nb. of Pubs	2.814^{**}	1.968^{**}	2.195^{**}
Top 0.1‰ of the Citation Distribution	(0.533)	(0.401)	(0.424)
Number of Applicants	3,075	3,075	3,075

Note: Each cell contains an estimate for the treatment effect in a separate regression. All estimates stem from Poisson regressions. The dependent variables are listed in the left-most column. All models incorporate a full suite of medical school graduation year effects as well as an indicator variable for holding a PhD degree at the time of application.
Exponentiated coefficients are presented; subtracting 1 yields magnitudes interpretable as elasticities. For example, the estimate in the bottom cell of the first column imply that attendees publish 100×(2.814-1)=181.4% more articles in the top 0.1‰ of the citation distribution during the independent phase of their career, relative to non-attendees; the effect is highly statistically significant. Columns 2 and 3 perform inverse probability of treatment and censoring weighted estimation; the corresponding estimates can be interpreted as the ATE/ATET of NIH training, under the assumption of unconfoundedness. Robust errors in parentheses ($^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$). *Sources*: ATP Index Cards, PubMed, Web of Science.

Table 6: Research style

	X-Sect.	Lasso Weights	
	Naive	ATE	ATET
Career Nb. of Pubs	1.609^{**}	1.478^{**}	1.505^{**}
	(0.117)	(0.112)	(0.131)
Basic Science Articles	2.787^{**}	2.197^{**}	2.184^{**}
	(0.320)	(0.291)	(0.348)
Translational Medicine Articles	1.830^{**}	1.542^{**}	1.570^{**}
	(0.196)	(0.179)	(0.207)
Clinical Trial Articles	1.584^{**}	1.544^{**}	1.674^{**}
	(0.188)	(0.178)	(0.205)
Other Clinical Articles	1.056	1.121	1.132
	(0.092)	(0.108)	(0.121)
Articles Appearing in	2.545^{**}	2.048^{**}	2.188^{**}
Translational Journals	(0.424)	(0.357)	(0.457)
Inspires Translational Research	1.799^{**}	1.583^{**}	1.623^{**}
	(0.211)	(0.186)	(0.207)
Builds on Translational Research	1.692^{**}	1.595^{**}	1.728^{**}
	(0.213)	(0.197)	(0.224)
Articles Cited in Patents	2.138^{**}	1.767^{**}	1.800^{**}
	(0.226)	(0.215)	(0.262)
Number of Applicants	2,584	2,584	2,584

Note: Each cell contains an estimate for the treatment effect in a separate regression. All estimates stem from Poisson regressions. The dependent variables are listed in the left-most column. All models also include a full suite of medical school graduation year effects as well as an indicator variable for holding a PhD degree at the time of application. Exponentiated coefficients are presented; subtracting 1 yields magnitudes interpretable as elasticities. For example, the estimate in the cell at the bottom left imply that attendees publish 100×(2.138-1)=113.8% more articles cite by patents, relative to non-attendees; the effect is highly statistically significant. Columns 2 and 3 perform inverse probability of treatment and censoring weighted; the corresponding estimates can be interpreted as the ATE/ATET of NIH training, under the assumption of unconfoundedness. Robust errors in parentheses ($^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$). *Sources*: ATP Index Cards, PubMed.

Table 7a: Robustness analyses

	IPTC Lasso Weights				Zero Pre-ATP Pubs	
	No Winsoring	Winsoring, 95th pctl.	Winsoring, 90th pctl.	CEM	Top 10 Med Schools	Other Med Schools
Career Nb. of Pubs	1.639^{**} (0.128)	1.516^{**} (0.124)	1.457^{**} (0.123)	1.921^{**} (0.219)	2.695^{**} (0.426)	1.675^{**} (0.235)
Log Pseudo-Likelihood	-152,478	-129,712	-113,235	-53,656	-47,084	-49,882
Number of Applicants	3,075	2,769	2,461	1,036	849	988

Note: Each cell contains an estimate for the treatment effect in a separate regression. All estimates stem from Poisson regressions. All models incorporate a full suite of medical school graduation year effects as well as an indicator variable for holding a PhD degree at the time of application. Exponentiated coefficients are presented; subtracting 1 yields magnitudes interpretable as elasticities. For example, the estimate in the first column imply that attendees publish 100×(1.639-1)=63.9% more articles during the independent phase of their career, relative to non-attendees. The first three columns vary the sample to reflect the winsorization of the inverse probability of treatment and censoring (IPTC) regression weights. In the fourth column, CEM refers to coarsened exact matching, a blocking technique to guarantee balance on a small set of covariates. The last two columns restrict sample to the set of applicants with no research experience prior to application, separately for those having graduated from elite and non-elite medical schools. Robust errors in parentheses ($^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$).

Table 7b: Robustness analyses

	Nb. of Pubs		$\sinh^{-1}$(Nb. of Pubs)	
	IPTC Lasso Weights	Double Lasso	IPTC Lasso Weights	Double Lasso
ATE	27.776^{**} (3.981)	25.776^{**} (4.000)	0.901^{**} (0.114)	0.868^{**} (0.085)
Number of Applicants	3,075	3,075	3,075	3,075

Note: Each cell contains an estimate for the average treatment effect in a separate regression. All estimates stem from OLS regressions. The dependent variable is either the number of post-training publications in levels (first pair of columns) or the inverse hyperbolic sine of the number of post-training publications (second pair of columns). The first and third columns perform inverse probability of treatment and censoring (IPTC) weighted estimation as in Table 4. The second and fourth column report an estimate of the average treatment effect using the "post-double-selection" lasso estimator due to Belloni et al. (2014). Robust errors in parentheses ($^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$).

Table 7c: Robustness analyses

	Oster's δ
Career Nb. of Pubs	1.743
Career Nb. of Pubs, Top 5% of the Cit. Distrib.	1.789
Career Nb. of Pubs, Top 1% of the Cit. Distrib.	1.767
Career Citations	1.748
Nb. of Patents	2.282
Career NIH Grants ($ 2015)	1.516

Note: The score reported corresponds to the δ parameter from Oster (2019), the ratio between the covariances of the outcome with observed and unobserved covariates, respectively. All outcomes are transformed using the inverse hyperbolic sine function, and δ is computed using OLS regression and the list of covariates selected by the pds-lasso estimator of Belloni et al. (2014), and chosen to produce an estimate of the treatment effect equal to zero. We follow Oster's recommendation of setting $R^{max} = 1.3 \times R^2$ from the fully saturated specification.

Appendix A
Data Set Construction

Our initial source of data was the original application index cards of those who were interviewed on campus for the ATP. These cards included the applicants full name, application date, medical school and graduation year, any internship or residency training arranged by the time of application, honor society membership, and the NIH institutes and ATP program tracks to which they applied.

To obtain training and career histories, we manually collected and integrated data from multiple sources. State medical licensure records and websites such as ZocDoc, Doximity, and U.S. News and World Report Health frequently contained information on residency and fellowship training, current medical practice affiliations, and the dates holding an active medical license in each state. When available, we leveraged LinkedIn; professional profiles and CVs on their current employer's website; and Gale's Who's Who biographical profiles. For the 277 individuals who had passed away by 2017, we reviewed any published obituaries. Any deaths were also confirmed using the Social Security Death Index. For those with business or leadership roles, we reviewed executive biographies at resources such as Bloomberg, Crunchbase, or SEC filings. Google was also useful in identifying additional sources of career information, such as newspaper articles and press releases, biographies from speeches given or awards won, and other such public records. We also leveraged affiliations on publications to document postgraduate training programs and jobs. This was supplemented with data from the American Medical Association (AMA) Physician Master File, which includes year-by-year information on the geographic location and practice setting of clinical activities. To ascertain treatment status, participation in the ATP was verified with the biographical resources above as well as NIH telephone directories and internal human resource records. We took great pains to be very conservative in any cases where judgement was required. Together, these allowed us to identify the location and dates of any post-graduate training, time of career independence, first job after independence, current job or last job prior to retirement, and year of retirement or death (if applicable). Only 13 applicants were lost to follow up using these combined sources.

Potential publications, patents, and NIH grants for each individual were identified after careful name-disambiguation and then manually verified. Verification capitalized on the biographical information collected above, including institutional affiliation, geographic location, and research interests.

The richness of our individual career histories lets us measure participation in research independently of the applicants' employers. For instance, the career of many of our applicants unfolds within academic medical centers in purely clinical positions where there is no expectation of publication. In contrast, other applicants work in industry or other non-academic institutions and yet amass a respectable publication record in the context of non-traditional research careers. The AMA Physician Master File distinguishes between individuals engaged in medical research, medical teaching, or direct patient care. For those in academic medicine, we also relied on the titles of positions held, which may characterize faculty as being on a "clinical track" (frequently with no expectation of research performance) or a research track. Finally, in the small number of cases where no explicit mention of titles and occupation could be found, we used the existence of publications to classify positions as research or clinically based.

Appendix B
Robustness Checks & Ancillary Results

Peer Coauthorship. A potentially important mechanism to explain the treatment effect of ATP attendance is peer effects. NIH ATPs could avail themselves to a much broader peer community than in the typical lab or clinical fellowship in an elite university where non-attendees might have completed training, and the oral histories are replete with mentions of the concentration in talent at the NIH brought about by the draft (see Appendix E). ATP attendees may have pushed their colleagues to work harder, hold higher internal standards of excellence or different approaches to problem solving, or helped instill in them values such as the inherent worth of translational investigations. Unfortunately, while we can observe the cohort of each ATP trainee, we cannot do the same for the non-attendee controls. Instead, we focus on a particular type of peer effects, that is driven by coauthorship.

For each ATP applicant, we identify their set of peer coauthors during training, and then observe persistence of coauthorship with these individuals during the independent phase of the career, symmetrically for both ATP attendees and non-attendees. This approach relies on Author-ity for author-name disambiguation to identify the career publications of applicants' coauthors (Torvik et al., 2005; Torvik and Smalheiser, 2009). To ensure our measure of coauthorship from peers during training does not include established research mentors, we exclude coauthors who published their first paper over 15 years before the focal collaboration (which allows for the sometimes extended training of any peers who have an MD/PhD and remain in fellowship or postdoctoral training) as well as those who have published at least 3 papers together with an ATP applicant on which they were last author for at least 75% of their shared papers during the applicants training. We apply these measures to the 2,096 physicians (1,549 attendees and 547 non-attendees) who have both at least one publication during training and one publication during career independence. We limited our analysis to publications that appeared before or in 2009, the years for which author-name disambiguation data is available in Author-ity.

We first look at descriptive statistics (Table B8a). While the absolute number of papers sharing a peer coauthor from training increases from an average of 5.8 to 8.6 for ATP attendees compared to non-attendees, the fraction of papers sharing a peer coauthor from training among all of their post-independence publications decreases from 13.3% to 12.4%. We also measure coauthorship for the subset of ATP attendees who take their first job following career independence as an investigator at the NIH. Not surprisingly, the percentage of papers with a peer coauthor is higher among this population (16.3%).

We estimate the total number of coauthored papers using a Poisson model with an offset for the total number of publications after career independence (Table B8b). This may be interpreted as showing the change in the fraction of the total publications with a peer coauthor from the training period. The coefficients are small in magnitude and not statistically significant when comparing all ATP attendees to non-attendees. The subset of ATP attendees who take their first job after career independence at the NIH are more likely to publish during career independence with peer coauthors from training; this effect is statistically significant. This likely reflects both collaboration with former mentors as well as with peers who themselves become principal investigators at the NIH. While we do not find evidence supporting a large role for coauthorship-driven peer effects, this does not preclude an important role for other types of peer effects.

Specialization. Increased specialization is one potential mechanism contributing to increased ATP productivity. We take advantage of MeSH terms, a hierarchical controlled vocabulary thesaurus maintained by the National Library of Medicine, to characterize the topics associated with each ATP applicants publications throughout their career. Importantly, authors are not involved in selecting MeSH terms for their publications; instead, MeSH terms are selected by professional indexers employed by the National Library of Medicine.

We employ two indices to measure topic diversity. The first follows the index proposed by Nagle and Teodoridis (2020) and Teodoridis et al. (2019): one minus the Herfindahl index over all MeSH keywords in the researcher's independent career (hereafter diversification index). The second is the Ellison-Glaeser concentration index (EG index), initially developed to study the geographic concentration of U.S. manufacturing industries (Ellison and Glaeser, 1997). This measure alleviates two problems that affect simpler measures of density like the Herfindahl index. First, many of the ATP applicants in our sample tend to have a small number of publications; this is often case for those ATP applicants who exit research early in their career for primarily clinical positions. This makes it hard to tell "true" specialization—as a consequence of the agenda chosen by researchers—from lack of actual variation due to insufficient opportunity to observe enough publications, and consequently, enough MeSH terms. Similarly, as scientific careers progress, researchers may naturally explore different topics as fields advance and previous motivating research questions become answered. Figure B4a, which plots each ATP applicant's diversification index against his post-independence publications, show this is indeed a problem in our data. Moreover, this problem does not cut symmetrically across attendees and non-attendees: ATP attendance leads to a large increase in the number of publications and length of time in research jobs (Tables 3 and 4). We might therefore expect to mechanically observe more diversity of topics pursued by ATP attendees relative to non-attendees, but this might mostly reflect the less productive and shorter research careers typically experienced by non-attendees.

There is a second problem associated with Herfindahl-based measures of diversification. ATP attendance shifted attendees towards basic science and translational research; non-attendees have a comparatively higher percentage of their publications in clinical research. The number and density of available MeSH terms vary across fields. The observed effect could simply reflect shifting to basic or translational science rather than reflecting true differences in the degree of specialization.

As a result, we would like to assess the degree to which researchers exhibit "excess" focus relative to some baseline. In our case, the most natural baseline is the distribution of MeSH keywords in the sample as a whole. A high value of the EG index will mean that a researcher is more focused than would have been expected if its keywords had been distributed in a similar way as the sample-wide distribution of keywords. The EG index appropriately adjusts the Herfindahl index to reflect these deviations from the benchmark. Because we measure diversification rather than focus, our outcome of interest is one minus the EG-index, which we refer to as the EG-index of diversification, in a slight abuse of nomenclature. Figure B4b shows the EG index entails much less dramatic variation in the average degree of diversification as one moves from the set of researchers with lower levels of publications to highly productive researchers

We regress these outcome variables on the treatment indicator using the same econometric framework as in the main body of the manuscript, with the small caveat that we present OLS results since the EG index can take on negative values and the diversification index is bounded between zero and one (Table B9). Consistent with the conjectures above, we find that attendees pursue a more diverse set of topics over the career, relative to non-attendee controls, when measuring diversity using the diversification index. This effect disappears, however, when adopting the EG index-based measure, which we prefer for the reasons mentioned above. In summary, it does not appear that program attendance shifted intellectual specialization patterns using these metrics. As explained in the main body of the manuscript, however, we should not conclude from these results that the program did not affect the direction of the publications produced by attendees, relative to those of non-attendees (cf. Section 4.4 and Table 6).

Isolating the effect of informative censoring. We know from the results in Table 3 that the program shifts physicians from the clinical sector (where publication is considered at best a hobby) to the research sector (where publications and grants are absolutely key to career success). The large intensive margin magnitudes documented in Tables 4 and 5 could reflect, at least in part, the choice or opportunity to select into a position that affords the possibility of participating in idea-producing activities. To gauge whether this is the case, we could re-estimate the models corresponding to the outcomes in Table 4 on the subsample of physicians who begin their careers as researchers. However, since an initial placement in a research position lies on the causal pathway between training and research output, the estimates on the restricted sample cannot be given a causal interpretation.

Instead, we choose to analyze the effect of treatment using weights that adjust the naive estimates for selection into the program, but do not adjust them for selection into research careers (i.e., the corresponding specifications use IPT weights rather than IPTC weights [cf. Appendix E, eqns. (E.3.1) and (E.4.1)]). Appendix Table B10 reports these results. The magnitudes are always higher when using IPT weights instead of IPTC weights, but the differences between the two is not itself very large. Non-attendees publish less than attendees not simply by virtue of the fact the former are much less likely to be researchers. Rather, the effects on output reflect impacts at both the intensive and extensive margins.

Treatment effect dynamics. To examine how the impact of training varies over the course of an attendee's career, we first plot the raw average annual publications relative to the year of career independence for attendees and non-attendees (Figure B5a). Recall that typically physicians applied to the ATP in their final year of medical school and attended after completing the first year of residency, with an average length of about six years between medical school graduation and career independence. We observe a large difference in annual publications between treated and control physicians upon career independence, and this difference is sustained throughout the course of their career. Using the year of medical school graduation as a reference point since the length of post-graduate training is highly variable, we see that this divergence occurs, on average, primarily in years 3-5 after medical school graduation, corresponding to the typical years of ATP attendance (not shown). To more rigorously assess when the treatment effect occurs and how it varies over the course of careers, we used regression analysis with inverse probability of treatment and censoring weights (Figure B5b). Incorporating selection and censoring does change the patterns above. Focusing on publications as the key outcome, we do find the effect of the program is sustained over the entire career. While the U-shape visible in the raw comparison can also be discerned in regression models, it is more shallow. The coefficients corresponding to the different career time periods, while all significantly positive and different from zero, are not statistically different from one another.

Imprinting during training. While Table 6 presented results on the influence of NIH training on the direction of research pursued in the independent phase of a research career, Appendix Table B11 focuses on providing direct evidence of imprinting during training. To do so, the publication outcomes include only articles that appeared after one year from medical school graduation up to one year after the start of the independent career (to allow for publication lags). Unsurprisingly, NIH trainees publish more than non-attendee controls in training. But the type of publication published also differs markedly from that exhibited by control trainees. Attendees' publications are much less likely to fall in the "other clinical" category, for instance.[i]

Heterogeneity across time. Appendix Table B12 examines whether the program's effects varied in magnitude over the time period considered in this study. Recall from Figure 1 that we have only a small number of controls available in the early part of the sample (1965-1969). It is also possible that the incentives to apply (and to attend if selected) decreased in the waning years of the Vietnam conflict and the impending end of the draft (1973-1975). We find that program attendance impacted initial placement in research-focused jobs regardless of time period. In contrast, effects at the intensive margins (e.g., post-training career publications) are lower and less precisely estimated in the latter part of the observation period. This

[i] In the subsamples for each track, we only include as controls trainees who applied unsuccessfully for that track.

attenuation might reflect a decrease in the quality of the applicant pool, but a more cogent explanation is that the progressive availability of high-quality research training outside the confined boundaries of NIH's Bethesda campus boosted the outcomes for non-attendees.

Heterogeneity by program track. Appendix Table B13 sheds light on whether the effects of program attendance differ according to the track (research associate, clinical associate, or staff associate) for which attendees were selected. Across a broad range of outcome variables, we do not find evidence of markedly different magnitudes for the research and clinical associate tracks, whereas the post-training record of staff associates appears slightly less distinguished. Similarly, we find little evidence of significant differences in the style-composition of the research portfolio for scientists in these tracks (once again this is less true for the staff associate track). While perhaps surprising, it is important to note that research associates and clinical associates ultimately often worked in the same labs, and the distinction between research and clinical time was not always clear-cut in practice.

References

Ellison, Glenn, and Edward L. Glaeser. 1997. "Geographic Concentration in U.S. Manufacturing Industries: A Dartboard Approach." *Journal of Political Economy* **105**(5):889-927.

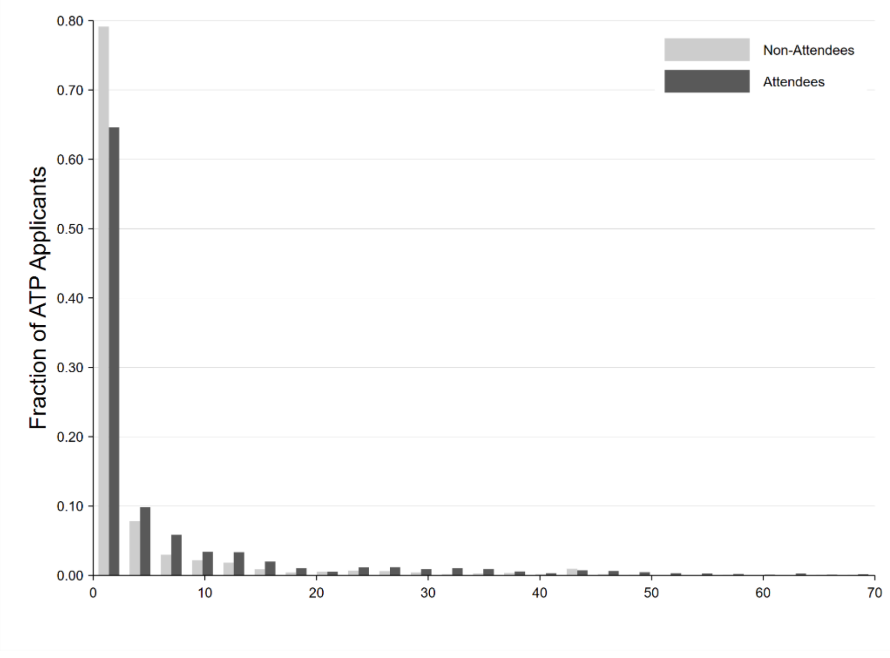
Nagle, Frank, and Florenta Teodoridis. 2019. "Jack of all trades and master of knowledge: The role of diversification in new distant knowledge integration." *Strategic Management Journal* **41**(1):55-85.

Teodoridis, Florenta, Michael Bikard, and Keyvan Vakili. 2018. "Creativity at the Knowledge Frontier: The Impact of Specialization in Fast- and Slow-paced Domains." *Administrative Science Quarterly* **64**(4):894-927.

Torvik, Vetle I., and Neil R. Smalheiser. 2009. "Author Name Disambiguation in MEDLINE." *ACM Transactions on Knowledge Discovery from Data* **3**(3):11.

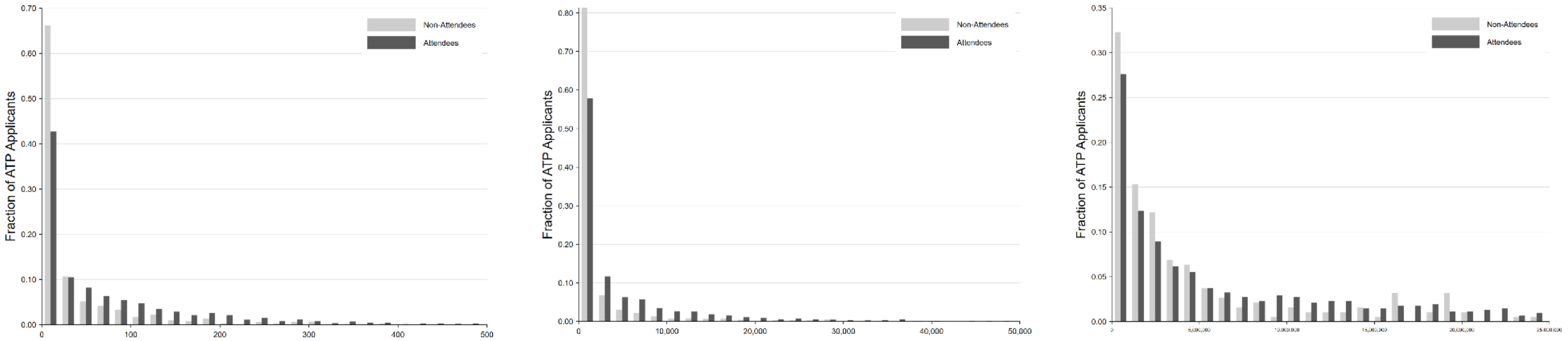
Torvik, Vetle I., Marc Weeber, Don R. Swanson, and Neil R. Smalheiser. 2005. "A probabilistic similarity metric for Medline records: a model for author name disambiguation." *Journal of the American Society for Information Science and Technology* **56**(2):140-158.

Figure B1. Pre-application publications



Note: Histogram for the number of original publications published up to the year of ATP application weighted by the journal impact factor of the publication outlet in which they appeared. Twenty-four outliers are omitted.

Figure B2. Career research outcomes



A. Career publications

Note: Histogram for the number of original journal publications over the entire post-training career. Twenty-five outliers with more than 500 post-training publications omitted. 86% of applicants publish one article or more after career independence (91% of attendees; 78% of non-attendees). *Sources*: ATP Index Cards and PubMed.

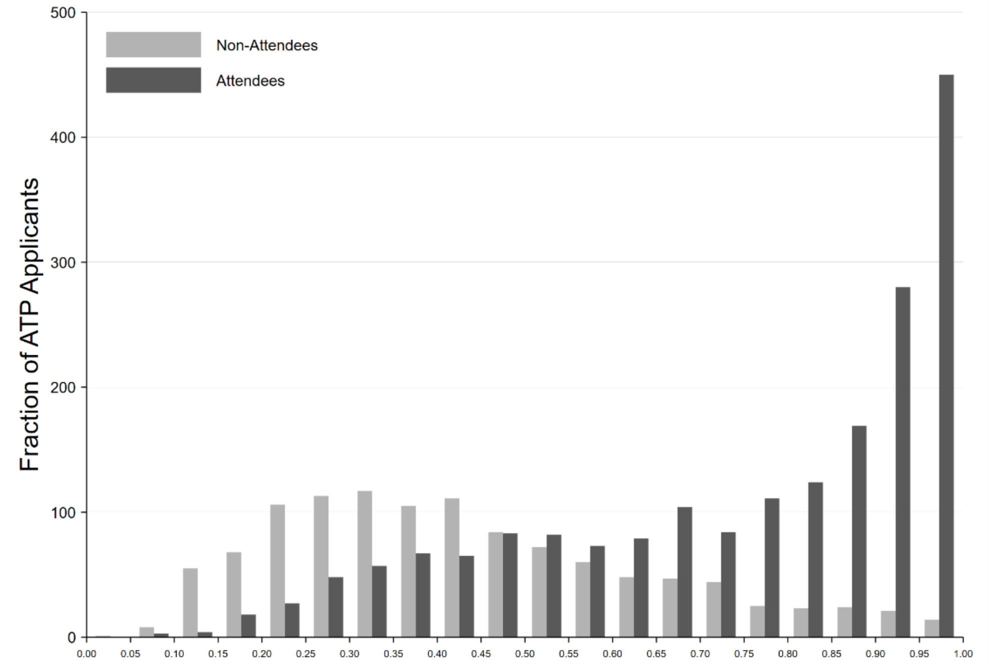
B. Career citations

Note: Histogram for the cumulative number of citations to original journal publications published over the entire post-training career. Twenty outliers with more than 50,000 citations omitted (excludes citations to publications as a student or trainee). *Sources*: ATP Index Cards, PubMed, and Web of Science.

C. Career NIH funding

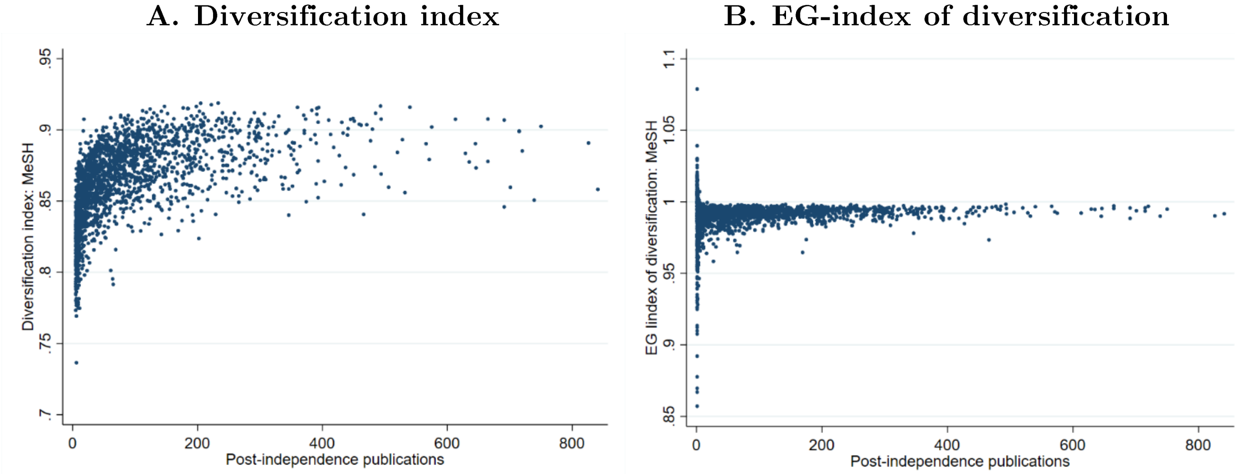
Note: Histogram for the cumulative NIH funding received over the entire post-training career (2015 dollars, deflated by the Biomedical R&D PPI). 1,087 fellows receive at least some NIH funding during their career. Two hundred and eighty four outliers with more than $25 mln. in career funding omitted. *Sources*: ATP Index Cards and NIH Compound Grant Applicant File (CGAF).

Figure B3. Predicted probability of selection for logit specification



Note: Predicted probabilities from the logit specification reported in Table 2, column (1b). The correlation coefficient between the logit weights and those derived from the lasso penalized logit procedure is 0.92.

Figure B4. Specialization and career research productivity



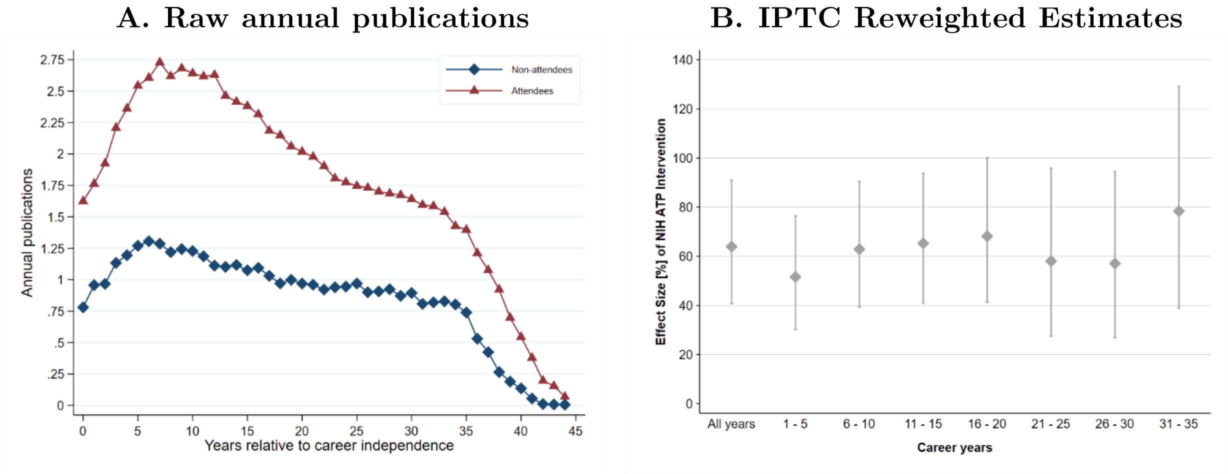
A. Diversification index

B. EG-index of diversification

Note: This figure plots one minus the Herfindahl index as the measure of diversity against each researcher's post-independence publications. Each data point represents a single ATP applicant with at least one post-independence publication (n=2,647). *Sources*: PubMed.

Note: This figure plots one minus the Ellison-Glaeser (EG) index as the measure of diversity against each researcher's post-independence publications. Each data point represents a single ATP applicant with at least one post-independence publication (n=2,647). *Sources*: PubMed.

Figure B5. Treatment effect dynamics



A. Raw annual publications

B. IPTC Reweighted Estimates

Note: This figure plots the raw mean annual publication count for ATP applicants by attendance status relative to the year of career independence. *Sources*: ATP Index Cards and PubMed.

Note: Each estimate corresponds to the treatment effect in a separate Poisson regression model with an outcome of the number of publications during the career period as specified along the x-axis. All models incorporate a full suite of medical school graduation year effects as well as an indicator variable for holding a PhD degree at the time of application. Inverse probability of treatment weights are used calculated from lasso specifications similar to that appearing in Table 2; the corresponding estimates can be interpreted as the ATE of NIH training, under the assumption of unconfoundedness. Exponentiated coefficients are presented. The 95% confidence intervals are computed using robust standard errors. *Sources*: ATP Index Cards and PubMed.

Table B1. Most common medical schools attended by applicants

Medical School	Non-Attendees	Attendees	Total
Harvard Medical School	86 (7.50)	274 (14.20)	360 (11.71)
Johns Hopkins University School of Medicine	54 (4.71)	113 (5.86)	167 (5.43)
Columbia University College of Physicians & Surgeons	57 (4.97)	85 (4.41)	142 (4.62)
University of Pennsylvania School of Medicine	53 (4.62)	87 (4.51)	140 (4.55)
New York University School of Medicine	45 (3.93)	84 (4.35)	129 (4.20)
Yale University School of Medicine	52 (4.54)	77 (3.99)	129 (4.20)
Albert Einstein College of Medicine of Yeshiva University	52 (4.54)	63 (3.27)	115 (3.74)
Duke University School of Medicine	22 (1.92)	75 (3.89)	97 (3.15)
SUNY Downstate Medical Center College of Medicine	38 (3.32)	51 (2.64)	89 (2.89)
Cornell University Medical College	30 (2.62)	52 (2.70)	82 (2.67)
Total	489 (42.67)	961 (49.82)	1,450 (47.15)

Note: Column percentages in parentheses.

Table B2. Occupational breakdown, first position (post-training)

First Position	Non-Attendees	Attendees	Total
Academic Researcher	509 (44.42)	1,170 (60.65)	1,679 (54.60)
Academic Clinician	131 (11.43)	130 (6.74)	261 (8.49)
NIH Staff Scientist	15 (1.31)	161 (8.35)	176 (5.72)
Solo Clinical Practice	176 (15.36)	143 (7.41)	319 (10.37)
Group Clinical Practice	233 (20.33)	203 (10.52)	436 (14.18)
Hospital Clinical Practice	73 (6.37)	95 (4.92)	168 (5.46)
Industry	2 (0.17)	7 (0.36)	9 (0.29)
Biopharma Consulting	1 (0.09)	1 (0.05)	2 (0.07)
Administrative Position	1 (0.09)	2 (0.10)	3 (0.10)
Health & Science Policy	2 (0.17)	15 (0.78)	17 (0.55)
Miscellaneous	3 (0.26)	2 (0.10)	5 (0.16)
Total	1,146 (100.00)	1,929 (100.00)	3,075 (100.00)
N	3,075		

Note: Column percentages in parentheses.

Table B3. Occupational breakdown, last position

Last Position	Non-Attendees	Attendees	Total
Academic Researcher	295 (25.75)	851 (44.12)	1,146 (37.30)
Academic Clinician	134 (11.69)	168 (8.71)	302 (9.82)
NIH Staff Scientist	8 (0.70)	34 (1.76)	42 (1.37)
Solo Clinical Practice	209 (18.24)	222 (11.51)	431 (14.02)
Group Clinical Practice	317 (27.66)	344 (17.83)	661 (21.50)
Hospital Clinical Practice	93 (8.12)	116 (6.01)	209 (6.80)
Industry	24 (2.09)	79 (4.10)	103 (3.35)
Biopharma Consulting	17 (1.48)	38 (1.97)	55 (1.79)
Administrative Position	24 (2.09)	50 (2.59)	74 (2.41)
Health & Science Policy	15 (1.31)	18 (0.93)	33 (1.07)
Miscellaneous	10 (0.87)	9 (0.47)	19 (0.62)
Total	1,146 (100.00)	1,929 (100.00)	3,075 (100.00)
N	3,075		

Note: Column percentages in parentheses.

Table B4a. Covariate balance using logit weights

	Unweighted sample			Logit IPTW Reweighting		
	Non-attendees	Attendees	*t*-stat	Non-attendees	Attendees	*t*-stat
Pre-ATP JIF-Weighted Publications	3.288 (0.292)	6.595 (0.330)	6.835	6.349 (1.093)	4.587 (0.636)	1.540
NIH Grants for Applicant's Medical School	170.323 (3.792)	207.006 (3.438)	6.879	188.220 (8.776)	169.964 (22.130)	0.815
NIH Grants for Applicant's Internship Hospital	90.915 (2.590)	97.153 (1.882)	1.986	83.782 (4.267)	82.556 (10.984)	0.105
PhD	0.013 (0.003)	0.036 (0.004)	3.738	0.030 (0.011)	0.024 (0.004)	0.585
No Internship	0.002 (0.001)	0.009 (0.002)	2.532	0.005 (0.005)	0.006 (0.002)	0.155
Applies more than once	0.027 (0.005)	0.028 (0.004)	0.154	0.041 (0.010)	0.025 (0.005)	1.604
AΩA Honor Medical Society	0.257 (0.013)	0.383 (0.011)	7.162	0.328 (0.028)	0.294 (0.039)	0.706
Attended Harvard Medical School	0.075 (0.008)	0.142 (0.008)	5.614	0.106 (0.021)	0.110 (0.016)	0.130
Attended Johns Hopkins School of Medicine	0.047 (0.006)	0.059 (0.005)	1.356	0.041 (0.008)	0.050 (0.008)	0.797
Attended Columbia University	0.050 (0.006)	0.044 (0.005)	0.725	0.052 (0.011)	0.037 (0.007)	1.191

Note: Means, standard errors, and t-statistics are reported; reweighting is performed using average treatment effect inverse probability of treatment weights. *t*-statistics are calculated using IPTW-weighted OLS regression of the variable of interest on an indicator variable for ATP attendance. Harvard, Johns Hopkins, and Columbia are the three most common medical schools attended in the sample. For NIH grants, original amounts were deflated using the Biomedical R&D Producer Price Index (2015 dollars) and presented in units of millions of dollars. JIF—journal impact factor. *Sources*: ATP Index Cards, PubMed, CGAF.

Table B4b. Research and training outcomes using logit weights

	X-Sect.	Lasso Weights		Logit Weights	
	Naive	ATE	ATET	ATE	ATET
Logit Estimates					
Researcher First Job	0.212**	0.170**	0.153**	0.179**	0.161**
	(0.018)	(0.022)	(0.026)	(0.026)	(0.033)
Ends Career as Researcher	0.216**	0.192**	0.175**	0.140**	0.175**
	(0.019)	(0.030)	(0.025)	(0.022)	(0.030)
Poisson Estimates					
Career Nb. of Pubs	1.922**	1.639**	1.668**	1.698**	1.752**
	(0.146)	(0.128)	(0.152)	(0.150)	(0.196)
Career citations	2.316**	1.775**	1.884**	1.753**	1.954**
	(0.227)	(0.190)	(0.220)	(0.217)	(0.293)
Nb. of Patents	2.515**	1.635*	1.559†	1.252	1.290
	(0.521)	(0.368)	(0.406)	(0.306)	(0.412)
Career NIH R01 Grants ($ 2015)	2.285**	1.668**	1.774**	1.442†	1.713*
	(0.352)	(0.296)	(0.340)	(0.307)	(0.429)
Nb. NIH-R01-funded trainees	2.410**	1.621*	1.689*	1.348	1.563
	(0.365)	(0.349)	(0.429)	(0.396)	(0.532)
Number of Applicants	3,075	3,075	3,075		

Note: Each cell contains an estimate for the treatment effect in a separate regression. The dependent variables are listed in the left-most column. All models incorporate a full suite of medical school graduation year effects as well as an indicator variable for holding a PhD degree at the time of application. In the first two rows, the estimates stem from logistic regressions. The marginal effects for the treatment indicator are reported. For instance, the coefficient in the first row of the first column implies that attendees are 21.2% more likely than non-attendees to be initially placed in a research position after completing their training. In the remaining rows, the estimates stem from Poisson regressions. Exponentiated coefficients are presented; subtracting 1 yields magnitudes interpretable as elasticities. For example, the estimate in the first column of the third row imply that attendees publish 100×(1.922-1)=92.2% more original publications after career independence, relative to non-attendees. The last four columns perform inverse probability of treatment weighted estimation for first career position and inverse probability of treatment and censoring weighted estimation for all other outcomes; the corresponding estimates can be interpreted as the ATE/ATET of NIH training, under the assumption of unconfoundedness. Robust errors in parentheses ($^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$). *Sources*: ATP Index Cards, PubMed, Web of Science, CGAF, USPTO.

Table B5. Exclusion of controls not receiving research training

	Attendees vs. non-attendees		Attendees vs. limited non-attendee sample	
	X-Sect. Naive	Lasso weights ATE	X-Sect. Naive	Lasso weights ATE
Logit Estimates				
Researcher First Job	0.160**	0.111**	0.176**	0.139**
	(0.018)	(0.021)	(0.019)	(0.023)
Ends Career as Researcher	0.146**	0.131**	0.188**	0.155**
	(0.020)	(0.032)	(0.021)	(0.032)
Poisson Estimates				
Career Nb. of Pubs	1.922**	1.639**	1.739**	1.483**
	(0.146)	(0.128)	(0.133)	(0.119)
Observations	3,075	3,075	2,857	2,857

Note: Each cell contains an estimate for the treatment effect in a separate regression. The dependent variables are listed in the left-most column. All models incorporate a full suite of medical school graduation year effects as well as an indicator variable for holding a PhD degree at the time of application. In the first two rows, the estimates stem from logistic regressions. The marginal effects for the treatment indicator are reported. For instance, the coefficient in the first row of the first column implies that attendees are 16.0% more likely than non-attendees to be initially placed in a research position after completing their training. In the remaining row, the estimate stems from Poisson regressions. Exponentiated coefficients are presented; subtracting 1 yields magnitudes interpretable as elasticities. For example, the estimate in the first column of the third row imply that attendees publish 100×(1.922-1)=92.2% more original publications after career independence, relative to non-attendees. Columns two and four perform inverse probability of treatment (row 1) and inverse probability of treatment and censoring (rows 2 and 3); the corresponding estimates can be interpreted as the ATE/ATET of NIH training, under the assumption of unconfoundedness. Robust errors in parentheses ($^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$). *Sources*: ATP Index Cards, PubMed.

Table B6. Intensity of Treatment Effects (dose-response relationship)

	Publications	Citations	NIH Funding	Patents	Academic First Job	Academic Last Job	Research First Job	Research Last Job
One year of NIH training	1.115	0.812	0.202*	0.349	1.071	1.072	1.155	1.126
	(0.438)	(0.450)	(0.148)	(0.324)	(0.152)	(0.138)	(0.172)	(0.136)
Two years of NIH training	1.486**	1.415**	1.158	1.200	1.099**	1.163**	1.165**	1.239**
	(0.118)	(0.158)	(0.342)	(0.329)	(0.025)	(0.038)	(0.030)	(0.036)
Three years of NIH training	2.063**	2.529**	3.654**	2.285**	1.187**	1.195**	1.256**	1.242**
	(0.211)	(0.322)	(1.492)	(0.550)	(0.037)	(0.043)	(0.040)	(0.044)
Log Pseudo-Likelihood	-176,263	-14,078,663	-72,590,687,367	-10,356	-1,786	-2,364	-1,946	-2,319
Number of Applicants	3,075	3,075	3,075	3,075	3,075	3,075	3,075	3,075

Note: Each column reports estimates from a regression of the outcome listed in the header on four indicator variables corresponding to different intensities of treatment, as well as a full suite of medical school graduation year effects and an indicator variable for holding a PhD degree at the time of application. No NIH training is the omitted category of treatment intensity. The estimates stem from Poisson regressions (first four columns) and logit regressions (last four columns). For the Poisson estimates, exponentiated coefficients are presented; subtracting 1 yields magnitudes interpretable as elasticities. For the logit regressions, marginal effects are reported. Each observation is weighted by its inverse probability of treatment and censoring (columns 1-4, 6, and 8) or inverse probability of treatment (columns 5 and 7), as computed from a separate ordered logit specification very similar to that appearing in Table 2. The corresponding estimates can be interpreted as the ATE of NIH training, under the assumption of unconfoundedness. Robust errors in parentheses ($^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$). *Sources*: ATP Index Cards, PubMed, Web of Science, CGAF, USPTO, AMA Physician Masterfile.

Table B7a. Descriptive statistics: Mentor outperformance

	Non-attendees		Attendees	
	Mean	Std. Dev.	Mean	Std. Dev.
Peak citation percentile achieved in training (%ile)	80.26	23.25	91.36	13.05
Nb. of publications exceeding peak citation percentile achieved in training	1.83	5.69	2.76	9.15
Observations	728		1,782	

Note: Sample limited to the 2,510 (1,782 attendees and 728 non-attendees) with at least one publication during training included in Web of Science. *Sources*: ATP Index Cards, PubMed, Web of Science.

Table B7b. Mentor outperformance

		Lasso Weights
	Naive	ATE
Nb. of publications exceeding peak citation percentile achieved in training	1.368^{*} (0.211)	$1.288^{\dagger}$ (0.182)
Number of Applicants	2,510	2,510

Note: Each cell contains an estimate for the treatment effect in a separate regression. All estimates stem from Poisson regressions with the dependent variable listed in the left-most column. All models incorporate a full suite of medical school graduation year effects as well as an indicator variable for holding a PhD degree at the time of application. Exponentiated coefficients are presented; subtracting 1 yields magnitudes interpretable as elasticities. For example, the estimate in the first column implies that attendees publish 100×(1.368-1)=36.8% more publications exceed the peak citation percentile achieved during training, relative to non-attendees. The second column performs inverse probability of treatment and censoring weighting; the corresponding estimates can be interpreted as the ATE of NIH training, under the assumption of unconfoundedness. The sample limited to the 2,510 (1,782 attendees and 728 non-attendees) with at least one publication during training included in Web of Science. Robust errors in parentheses ($^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$). *Sources*: ATP Index Cards, PubMed, Web of Science.

Table B8a. Descriptive statistics: Coauthorship peer effects

	Non-attendees		All attendees		Attendees staying at NIH	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Nb. of publications sharing peer coauthor from training	5.8	14.8	8.6	24.4	14.2	26.0
Nb. of publications not sharing peer coauthor from training	49.7	75.5	71.2	88.0	78.4	78.3
Fraction of publications with peer coauthor from training	0.133	0.238	0.124	0.218	0.163	0.225
Observations	547		1,549		143	

Note: *Sources*: ATP Index Cards, PubMed, Author-ity.

Table B8b. Coauthorship peer effects

	Attendees vs. non-attendees	Attendees staying at NIH vs. non-attendees
Publications sharing peer coauthor from training	1.048 (0.144)	1.762^{**} (0.346)
Publications not sharing peer coauthor from training	0.994 (0.016)	0.908^{*} (0.038)
Number of Applicants	2,096	690

Note: Each cell contains an estimate for the treatment effect in a separate regression. The dependent variables measuring the number of post-independence publications either sharing or not sharing a coauthor from an individual's training period. All models incorporate a full suite of medical school graduation year effects. Inverse probability of treatment weights are used calculated from lasso specifications similar to that appearing in Table 2; the corresponding estimates can be interpreted as the ATE of NIH training, under the assumption of unconfoundedness. All estimates stem from Poisson regressions with an offset for the total number of post-independence publications. Exponentiated coefficients are presented; subtracting 1 yields magnitudes interpretable as elasticities. For example, the estimate in the first row of column 1 implies that attendees publish 100×(1.048-1)=4.8% more publications during career independence with a coauthor from training, relative to non-attendees; this result is not statistically significant. N = 2,096 (1,549 attendees and 547 non-attendees) for column 1 and 718 (143 attendees and 547 non-attendees) for column 2. ($^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$). *Sources*: ATP Index Cards, PubMed, Author-ity.

Table B9. Researcher specialization

	Diversification index		EG-index of diversification	
		Lasso Weights		Lasso Weights
	Naive	ATE	Naive	ATE
ATP Attendee	0.0303**	0.0239**	0.0004	0.0009
	(0.0031)	(0.0033)	(0.0005)	(0.0006)
R^2	0.0765	0.0473	0.0100	0.0112
Mean Index Score	0.8357	0.8385	0.9903	0.9902
Number of Applicants	2,647	2,647	2,647	2,647

Note: Each cell contains an estimate for the treatment effect in a separate regression. The dependent variables is the diversification index (columns one and two) and EG-index of diversification (columns three and four). All models stem from OLS regression and incorporate a full suite of medical school graduation year effects as well as an indicator variable for holding a PhD degree at the time of application. The second and fourth column performs inverse probability of treatment and censoring weighting; the corresponding estimates can be interpreted as the ATE of NIH training, under the assumption of unconfoundedness. The sample is limited to those APT applicants with at least 1 publication after career independence with associated MeSH terms. Robust errors in parentheses ($^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$).

Table B10. Research and training outcomes by weighting procedure

	Naive Estimates X-Sect.	Reweighted Estimates IPTW	Reweighted Estimates IPCW	Reweighted Estimates IPTCW
Logit Estimates				
Researcher First Job	0.212^{**} (0.018)	0.170^{**} (0.022)		
Ends Career as Researcher	0.216^{**} (0.019)	0.192^{**} (0.024)	0.180^{**} (0.021)	0.192^{**} (0.030)
Poisson Estimates				
Career Nb. of Pubs	1.922^{**} (0.146)	1.694^{**} (0.134)	1.722^{**} (0.132)	1.639^{**} (0.128)
Career citations	2.316^{**} (0.227)	1.931^{**} (0.207)	2.012^{**} (0.206)	1.775^{**} (0.190)
Nb. of Patents	2.515^{**} (0.521)	1.883^{**} (0.415)	2.230^{**} (0.463)	1.635^{*} (0.368)
Career NIH R01 Grants ($ 2015)	2.285^{**} (0.352)	1.775^{**} (0.304)	2.075^{**} (0.322)	1.668^{**} (0.296)
Nb. NIH-R01-funded trainees	2.410^{**} (0.365)	1.707^{*} (0.371)	2.177^{**} (0.325)	1.621^{*} (0.349)
Number of Applicants	3,075	3,075	3,075	3,075

Note: Each cell contains an estimate for the treatment effect in a separate regression. The dependent variables are listed in the left-most column. All models incorporate a full suite of medical school graduation year effects as well as an indicator variable for holding a PhD degree at the time of application. In the first two rows, the estimates stem from logistic regressions. The marginal effects for the treatment indicator are reported. For instance, the coefficient in the first row of the first column implies that attendees are 21.2% more likely than non-attendees to be initially placed in a research position after completing their training. In the remaining rows, the estimates stem from Poisson regressions. Exponentiated coefficients are presented; subtracting 1 yields magnitudes interpretable as elasticities. For example, the estimate in the first column of the third row imply that attendees publish 100×1.922-1)=92.2% more original publications after career independence, relative to non-attendees. The first column corresponds to a naïve cross sectional estimate of the difference in outcomes for treated and control applicants, controlling for a handful of predetermined covariates. The second column corresponds to reweighted estimates using inverse probability of treatment weights, which adjust the effect for selection but ignore informative censoring. The third column corresponds to reweighted estimates using inverse probability of censoring weights, which adjust the effect for early exit from research but ignore selection concerns. The fourth column combines the two sets of weights, producing estimates robust to selection and informative censoring under the maintained assumption of unconfoundedness. All weights are average treatment effects calculating from a lasso model. Robust errors in parentheses ($^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$). *Sources*: ATP Index Cards, PubMed, Web of Science, CGAF, USPTO.

Table B11. Research outcomes and style during training, by program track

	All Associates		Research Assoc.		Clinical Assoc.		Staff Assoc.	
	Naive	ATE	Naive	ATE	Naive	ATE	Naive	ATE
Poisson estimates for full sample								
Nb. of Pubs, Training Period	2.405**	2.054**	2.426**	2.199**	2.525**	2.200**	2.587**	2.295**
	(0.142)	(0.163)	(0.179)	(0.168)	(0.161)	(0.167)	(0.278)	(0.263)
Number of Applicants	3,075	3,075	1,940	1,940	2,460	2,460	1,044	1,044
Poisson estimates for research style sample								
Nb. of Pubs, Training Period	1.516**	1.432**	1.520**	1.507**	1.562**	1.495**	1.572**	1.450**
	(0.076)	(0.077)	(0.092)	(0.103)	(0.082)	(0.087)	(0.144)	(0.146)
Basic Science Articles	2.881**	2.562**	2.721**	2.454**	2.791**	2.417**	2.696**	2.613**
	(0.253)	(0.242)	(0.308)	(0.310)	(0.271)	(0.261)	(0.451)	(0.471)
Translational Medicine Articles	1.834**	1.670**	1.820**	1.756**	2.041**	1.853**	1.719**	1.561**
	(0.136)	(0.136)	(0.179)	(0.190)	(0.162)	(0.168)	(0.248)	(0.230)
Clinical Trial Articles	1.707**	1.651*	1.662*	1.729*	2.052**	2.154**	1.805	1.666
	(0.341)	(0.331)	(0.375)	(0.376)	(0.374)	(0.417)	(0.657)	(0.569)
Other Clinical Articles	0.881†	0.900	0.797*	0.825*	0.925	0.945	0.860	0.795
	(0.065)	(0.079)	(0.070)	(0.073)	(0.073)	(0.077)	(0.109)	(0.127)
Inspires Translational Research	1.749**	1.569**	1.820**	1.727**	2.039**	1.899**	1.682**	1.739**
	(0.179)	(0.173)	(0.236)	(0.233)	(0.222)	(0.231)	(0.332)	(0.318)
Builds on Translational Research	1.734**	1.678*	1.793*	1.943**	2.028**	2.201**	1.584	1.571
	(0.353)	(0.348)	(0.441)	(0.475)	(0.404)	(0.475)	(0.602)	(0.568)
Number of Applicants	2,486	2,486	1,625	1,625	1,968	1,968	862	862

Note: Each cell contains an estimate for the treatment effect in a separate regression. All estimates stem from Poisson regressions. The dependent variables are listed in the left-most column. All models incorporate a full suite of medical school graduation year effects as well as an indicator variable for holding a PhD degree at the time of application. The second, fourth, sixth, and eighth columns perform inverse probability of treatment weighted estimation as computed from lasso specifications very similar to that appearing in Table 2; the corresponding estimates can be interpreted as the ATE of NIH training, under the assumption of unconfoundedness. Exponentiated coefficients are presented; subtracting 1 yields magnitudes interpretable as elasticities. For example, the estimate in the top cell of the first column imply that attendees publish 100×(2.405-1)=140.5% more publications during training relative to non-attendees; the effect is highly statistically significant. The first row uses the full sample, while the other rows limit the sample to those with at least one cited publication during training. The number of applicants includes all those known to have applied to the corresponding track within the Associate Training Program. Robust errors in parentheses ($^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$). *Sources*: ATP Index Cards, PubMed, Web of Science.

Table B12. Treatment effect by application year

	All years		Applied 1965-1969		Applied 1970-1972		Applied 1973-1975	
	Naive	ATE	Naive	ATE	Naive	ATE	Naive	ATE
Researcher First Job	0.212^{**}	0.170^{**}	0.239^{**}	0.175^{*}	0.219^{**}	0.173^{**}	0.173^{**}	0.135^{**}
	(0.018)	(0.022)	(0.050)	(0.070)	(0.022)	(0.026)	(0.040)	(0.042)
Career Nb. of Pubs	1.922^{**}	1.639^{**}	2.335^{**}	2.029^{**}	2.103^{**}	1.767^{**}	1.309	1.148
	(0.146)	(0.128)	(0.491)	(0.489)	(0.186)	(0.164)	(0.224)	(0.196)
Total Applicants	3,075	3,075	1,037	1,037	1,570	1,570	468	468
Attendees	1,929	1,929	959	959	690	690	280	280
Non-Attendees	1,146	1,146	78	78	880	880	188	188

Note: Each cell contains an estimate for the treatment effect in a separate regression. The dependent variables are listed in the left-most column. All models incorporate a full suite of medical school graduation year effects as well as an indicator variable for holding a PhD degree at the time of application. The second, fourth, sixth, and eighth columns perform inverse probability of treatment (row 1) or inverse probability of treatment and censoring (row 2) weighted estimation as computed from lasso specifications very similar to that appearing in Table 2; the corresponding estimates can be interpreted as the ATE of NIH training, under the assumption of unconfoundedness. On the first row, the estimates stem from logistic regressions. The marginal effects for the treatment indicator are reported. For instance, the coefficient in the second row of the first column implies that attendees are 21.2% more likely than non-attendees to be initially placed in academia after completing their training. On the second row, the estimates stem from Poisson regressions. Exponentiated coefficients are presented; subtracting 1 yields magnitudes interpretable as elasticities. For example, the estimate in the second row of the first column imply that attendees publish 100×(1.922-1)=92.2% more original publications after career independence, relative to non-attendees; the effect is highly statistically significant. The number of applicants during each time period are also presented. Robust errors in parentheses ($^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$). *Sources*: ATP Index Cards, PubMed.

Table B13. Research outcomes and style, by program track

	All Associates		Research Assoc.		Clinical Assoc.		Staff Assoc.	
	Naive	ATE	Naive	ATE	Naive	ATE	Naive	ATE
Logit Estimates								
Researcher First Job	0.212**	0.170**	0.195**	0.148**	0.193**	0.142**	0.225**	0.175**
	(0.018)	(0.022)	(0.023)	(0.025)	(0.021)	(0.025)	(0.032)	(0.035)
Number of Applicants	3075	3075	1,940	1,940	2,460	2,460	1,044	1,044
Poisson Estimates for full sample								
Career Nb. of Pubs	1.922**	1.639**	1.918**	1.511**	1.917**	1.637**	1.926**	1.593**
	(0.146)	(0.128)	(0.181)	(0.143)	(0.164)	(0.145)	(0.258)	(0.211)
Number of Applicants	3075	3075	1,940	1,940	2,460	2,460	1,044	1,044
Poisson Estimates for research style sample								
Career Nb. of Pubs	1.609**	1.478**	1.641**	1.430**	1.625**	1.480**	1.585**	1.433**
	(0.117)	(0.112)	(0.148)	(0.131)	(0.132)	(0.127)	(0.202)	(0.179)
Basic Science Articles	2.787**	2.197**	2.732**	2.449**	2.730**	2.319**	2.443**	2.112**
	(0.320)	(0.291)	(0.392)	(0.349)	(0.370)	(0.328)	(0.501)	(0.444)
Translational Medicine Articles	1.830**	1.542**	1.872**	1.627**	1.927**	1.548**	1.532*	1.380†
	(0.196)	(0.179)	(0.251)	(0.209)	(0.241)	(0.217)	(0.302)	(0.264)
Clinical Trial Articles	(0.188)	1.544**	1.543**	1.221	1.698**	1.587**	1.531†	1.295
	(0.188)	(0.178)	(0.251)	(0.234)	(0.229)	(0.222)	(0.340)	(0.288)
Other Clinical Articles	1.056	1.121	1.016	0.918	1.075	1.119	1.201	1.128
	(0.092)	(0.108)	(0.119)	(0.108)	(0.100)	(0.107)	(0.199)	(0.171)
Inspires Translational Research	1.799**	1.583**	1.875**	1.634**	1.992**	1.620**	1.514†	1.325
	(0.211)	(0.186)	(0.273)	(0.235)	(0.274)	(0.231)	(0.340)	(0.286)
Builds on Translational Research	1.692**	1.595**	1.675**	1.294	1.854**	1.662**	1.684*	1.367
	(0.213)	(0.197)	(0.270)	(0.256)	(0.264)	(0.246)	(0.353)	(0.306)
Number of Applicants	2,584	2,584	1,685	1,685	2,061	2,061	899	899

Note: Each cell contains an estimate for the treatment effect in a separate regression. The dependent variables are listed in the left-most column. All models incorporate a full suite of medical school graduation year effects as well as an indicator variable for holding a PhD degree at the time of application. The second, fourth, sixth, and eighth columns perform inverse probability of treatment (row 1) or inverse probability of treatment and censoring (row 2-9) weighted estimation as computed from lasso specifications very similar to that appearing in Table 2; the corresponding estimates can be interpreted as the ATE of NIH training, under the assumption of unconfoundedness. On the first row, the estimates stem from logistic regressions. The marginal effects for the treatment indicator are reported. For instance, the coefficient in the first row of the first column implies that attendees are 21.2% more likely than non-attendees to be initially placed in a research position after completing their training. The estimates of the other rows stem from Poisson regressions. Exponentiated coefficients are presented; subtracting 1 yields magnitudes interpretable as elasticities. For example, the estimate in the second row the first column imply that attendees publish 100×(1.922-1)=92.2% more publications after career independence, relative to non-attendees; the effect is highly statistically significant. The first two rows uses the full sample, while the other rows limit the sample to those with at least one cited publication during career independence. The number of applicants includes all those known to have applied to the corresponding track within the Associate Training Program. Robust errors in parentheses ($^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$). *Sources*: ATP Index Cards, PubMed, Web of Science.

Appendix C
Non-applicant Sample

Identification of non-applicants. To construct information on the characteristics of physicians who did not apply to the NIH ATP, we capitalized on the American Medical Association (AMA) Physician Masterfile. The AMA Physician Masterfile was established in 1906; records for United States medical graduates are established upon medical school enrollment. We obtained records only for those physicians whose last name and first initial matched one of the applicants in our sample. We limited our analysis to male M.D. graduates from 1965 to 1975 who attended a U.S. medical school. Age was calculated directly from birthdates; we excluded those with an age greater than 36 out of concern for erroneous recorded birthdates. The final sample was comprised of 10,738 physicians who did not interview for an NIH ATP position.

To identify publications by non-applicants during medical school attendance, we focused on a subset of this population with a name frequency of 1 or 2 in our AMA Masterfile data (188 non-applicants and 1,502 applicants). Non-applicants were matched against author-name disambiguated publications using Authority (Torvik et al., 2005; Torvik and Smalheiser, 2009). All potential matches to publications were manually verified. Paralleling the measurement of pre-application publications by applicants, we kept only those original articles published within 1 year of medical school graduation.

Selection and treatment effect of applying. A natural question in light of the non-applicant sample is to understand the relative magnitude of selection into applying for the program with its treatment effect. Additionally, unsuccessful applicants may benefit from the exercise of applying itself (Ayoubi et al., 2019). Unfortunately, our research design and data do not allow us to shed light on the magnitudes of selection or the treatment effect of applying to the program.

First, we stress that the set of "non-applicants" might in fact include physicians who applied to the ATP but did not pass the first selection screen. In 1963, the only year for which reliable data is available for the number of applicants, about 21% of eligible male medical students applied. While reliable data is not available on the number of applicants in other years, oral histories and anecdotes suggest the number was much higher, perhaps as high as 80% of eligible male medical students (see section 1 and 2.1 for further details). In light of this, our non-applicant sample might better be termed "non-applicants and first-round rejected applicants." Second, only limited career information is available for non-applicants; some factors known to be associated with selection into the ATP, such as being a member of elite medical school honor societies ($A\Omega A$), are unavailable. Because of this, any observed differences between non-applicants and applicants are likely to reflect, at least in part, residual differences in talent and research inclination that we cannot adequately filter out with the covariates available.

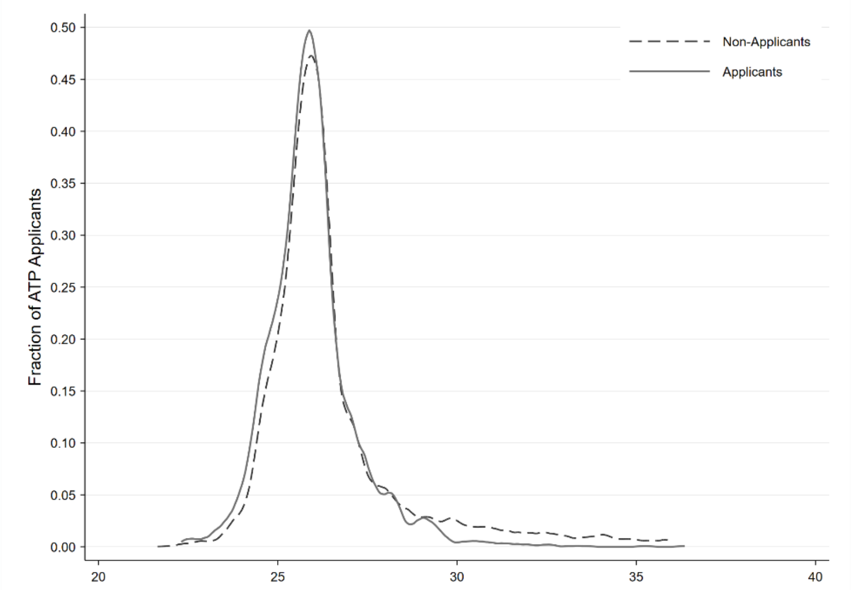
References

Ayoubi, Charles, Michele Pezzoni and Fabiana Visentin. 2019. "The important thing is not to win, it is to take part: What if scientists benefit from participating in research grant competitions?" *Research Policy* **48**(1):84 – 97.

Torvik, Vetle I., and Neil R. Smalheiser. 2009. "Author Name Disambiguation in MEDLINE." *ACM Transactions on Knowledge Discovery from Data* **3**(3):11.

Torvik, Vetle I., Marc Weeber, Don R. Swanson, and Neil R. Smalheiser. 2005. "A probabilistic similarity metric for Medline records: a model for author name disambiguation." *Journal of the American Society for Information Science and Technology* **56**(2):140-158.

Figure C1. Age at medical school graduation



Note: Kernel density of age at medical school graduation by application status. Non-applicants are male U.S. medical school graduates 1965-1975 identified from the American Medical Association Physician Masterfile (see Appendix B for details). N = 13,814 physicians (3,075 applicants and 10,738 non-applicants). *Sources*: ATP Index Cards, AMA Physician Masterfile.

Table C1. Descriptive statistics: Characteristics of non-applicants

	Mean	Median	Min.	Max.	Nb. of Obs.
Non-Applicants					
Age at medical school graduation	26.513	26.004	22	36	10,738
NIH Grants for Applicant's Med. School (×1,000)	103,459	77,069	0	639,320	10,738
Pre-ATP Nb. of Publications	0.314	0.000	0	5	188
ATP Interviewees					
Age at medical school graduation	25.952	25.845	22	36	3,075
NIH Grants for Applicant's Med. School (×1,000)	193,335	164,903	0	639,320	3,075
Pre-ATP Nb. of Publications	0.917	0.000	0	12	1,502

Note: Non-applicants are male U.S. medical school graduates between 1965 and 1975 identified from the American Medical Association Physician Masterfile. Pre-ATP publications are identified only for those with a name frequency of one or two (see Appendix B for details). For NIH grants, original amounts were deflated using the Biomedical R&D Producer Price Index (2015 dollars). *Sources*: ATP Index Cards, AMA Physician Masterfile, PubMed, CGAF.

Appendix D
Draft Lottery Subsample

We mentioned that physician applicants were *de facto* not eligible to participate in the draft lottery, having already made use of educational deferments. The exercise below counterfactually assumes that several cohorts of applicants in our sample entered the lotteries held by the U.S. Selective Service in 1969, 1970, and 1971. In total, 1,898 (61.72%) of the applicants were born between 1944 and 1952 and can therefore be counterfactually assigned a lottery number, based on their birth date. In the subsample of applicants thus "impacted" by the draft lottery, 978 (51.53%) have a number that was called, i.e., classified as available for military service.

Column 1 enters the same covariates into the specification as column 1a in Table 2, but estimates the model on the restricted sample of 1,898 "lottery-affected" applicants. Using this parsimonious model, column 2 shows that having one's lottery number called does not predict selection into the program, consistent with the premise that physicians were already on the required service list during the Vietnam era. Column 3 confirms this result using a lasso covariate selection procedure akin to that used in Table 2, column 1c (the draft lottery number indicator variable is constrained to appear in the specification).

Table D1. Modeling selection into the NIH ATP: Draft Lottery

	w/o Draft Covariates	w/ Draft Covariates	Saturated Model [Lasso]
	(1)	(2)	(3)
Draft Lottery Number Called		0.061	0.047
		(0.103)	(0.100)
Log(Pre-ATP Nb. of Publications)	0.294^{**}	0.294^{**}	0.302^{**}
	(0.078)	(0.078)	(0.079)
Ln(NIH Grants for Applicant's Medical School)	0.248^{*}	0.247^{*}	
	(0.109)	(0.109)	
Ln(NIH Grants for Applicant's Internship Hospital)	0.008	0.008	
	(0.011)	(0.011)	
PhD	0.744	0.747	1.087^{*}
	(0.494)	(0.496)	(0.453)
No Internship			2.770^{*}
			(1.143)
Applies more than once	0.119	0.117	0.433
	(0.305)	(0.307)	(0.282)
AΩA Honor Medical Society	0.634^{**}	0.634^{**}	0.623^{**}
	(0.121)	(0.121)	(0.115)
Constant	8.775^{**}	8.746^{**}	
	(2.178)	(2.179)	
Medical School Fixed Effects	No	No	Yes
Internship Hospitals Fixed Effects	No	No	Yes
Nb. of Non-zero Predictors			155
Nb. of Potential Predictors			369
χ^2 Test Statistic			52.22
Pseudo-R^2	0.144	0.144	
Log-likelihood	-1,125	-1,125	
Nb. of Applicants	1,898	1,898	1,898

Note: Estimates in columns 1a and 1b stem from logit specifications; the dependent variable is an indicator variable equal to one for attendees, zero for non-attendees. All models incorporate a full suite of medical school graduation year effects; a set of indicator variables for the applicant's age at the time of application; indicator variables for the number of distinct NIH component institutes that received the application; indicator variables for the number of tracks applied to within the Associate Training Program; indicator variables for the number of years between the application and the medical school graduation year; an indicator variable if the applicant applied more than once; an indicator variable for zero publications before application, and a series of indicator variables capturing if the applicant (1) intended to postpone his internship until after training, (2) intends to perform his internship abroad, (3) intends to intern in a hospital affiliated with the Veterans Affairs Administration, or (4) has missing information regarding his intended internship hospital. Estimates in column 1c correspond to the results of a cross-fit partialing-out lasso logit procedure with ten folds, as described in Chernozhukov et al. (2018). The specification includes all the covariates mentioned above, plus a full suite of medical school indicator variables and a full suite of internship hospitals indicator variables, for a total of 369 covariates, 155 of which the procedure selects for inclusion as control variables. The χ^2 test statistic (i.e., the Wald test of the hypothesis that the coefficients of the five variables of interest—for which inference is performed and are constrained to appear in the model—are jointly equal to zero) is equal to 52.22 ($p < 0.01$). Robust errors in parentheses ($^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$). *Sources*: ATP Index Cards, AMA Physician Masterfile, PubMed, CGAF, draft lottery numbers available at https://www.sss.gov/history-and-records/vietnam-lotteries/.

Appendix E
Background Institutional Detail

The Research Environment at the National Institutes of Health Intramural Campus. NIH traces its roots to 1887, when a one-room "Laboratory of Hygiene" was created within the Marine Hospital Service, a predecessor agency to the U.S. Public Health Service. This laboratory evolved into the Hygienic Laboratory, which moved to Washington, D.C. in 1891 and, with the Ransdell Act of 1930, became the National Institute of Health. NIH remained primarily an intramural effort until after World War II, although it collaborated with academic institutions during wartime to solve war-related health problems such as the need for large-scale production of penicillin and the need for new drugs for malaria treatment. In 1944, the Public Health Service Act authorized the Public Health Service to make grants to universities, laboratories, and hospitals for the conduct of research. The early success of the extramural funding component of the National Cancer Institute (which became part of NIH in 1947) laid the foundation for the concept of a health research agency relying on a mix of extramural and intramural research (NIH Office of History, 1982).

After the war, Vannevar Bush, director of the Office of Scientific Research and Development, outlined a program for postwar scientific research which affirmed the contributions of "remote and unexpected fields of medicine and the underlying sciences" in the progress against disease, and the benefits of cooperative endeavors with industry and academia (Cassell et al., 1994). Over the next few decades, Congress greatly increased funding to the NIH, and various institutes and centers within the NIH were created for specific research programs. The 1953 opening of the Clinical Center on the NIH campus added a new dimension to the intramural program—a large capability for patient-related research in close proximity to basic research laboratories—and brought to the campus a new complement of physicians as well as other professionals and staff needed to run a research hospital (Shapiro et al., 1988).

By the early 1960s, the NIH intramural program had become an elite research institution, a fact acknowledged by a committee report criticizing the rationale for its existence on the grounds that *". . . the government should not undertake the direct conduct of research activities that fit precisely into the pattern of scientific work that the universities or other non-government institutions are equipped to perform"* (Wooldridge Committee Report, 1965).[ii]

Oral histories from NIH staff are replete with claims attesting to the cutting edge research, deep expertise, and concentration of talent in biomedical research within the confines of the intramural campus that resulted in a rarefied environment.[iii] Alan Schechter, an ATP fellow and later Chief of the Molecular Medicine Branch at the NIH noted the NIH created much of the psychiatric research at the biochemical level, research which was *"almost non-existent anywhere else in the country, or throughout the world, before it was started at the [National Institute of Mental Health] in Bethesda"* (Schechter and Schechter, 1998). Similarly, Vincent DeVita, an ATP fellow who helped to develop the first successful combination chemotherapy program and held a series of leadership positions at the National Cancer Institute, Memorial Sloan Kettering Cancer Center, and Yale recalls that *"I was going to stay at Yale for a fellowship, as well, but after I got here I realized that what I was doing at the NIH was so much more advanced in the cancer field than what was going on here at Yale, that I just decided to go back"* (DeVita, 1997). Donald Fredrickson, the Director of NIH between 1975 and 1981, emphasized the breadth of expertise available, noting *"you had an expert in virtually everything you were working on right here on this campus,"* so often collaborating or learning new techniques was a simple walk down the hallway (Fredrickson, 1998).

[ii] Writing a letter to *Science* in defense of the intramural program, Alfred Gellhorn, a prominent physician-scientist at Columbia University noted the importance of research training for physicians, asserting that *". . . this enlargement of manpower in the health-research sciences would be justification enough for the intramural program"* (Gellhorn, 1965).

[iii] `https://history.nih.gov/archives/oral_histories.html`

In addition to its scientific expertise, the NIH at the time had several unique strengths. Its culture is often described as more similar to a university than to a government research lab. Unlike a university, however, investigators had fewer administrative and teaching responsibilities, allowing them to focus a greater proportion of their time on research. Resources were plentiful, and intramural funding protected investigators from the vagaries of the grant peer review process. The culture encouraged researchers to be independent and explore their own ideas wherever they led. Edward Scolnick, an ATP fellow and former head of research and development at Merck Research Laboratories noted that at the NIH, *"the obligation was that you shouldn't just turn out the papers. You really had to work on something that was truly meaningful"* (Scolnick, 1998).

The draft, and resulting remarkable concentration of talent in one location, also contributed to the research environment at the NIH. Harry Kimball, an ATP fellow and former President of the American Board of Internal Medicine, recalls, *"the very best people and trainees in the country came to the NIH and that of course led to an atmosphere and an environment which was truly remarkable... Make no mistake the draft concentrated a number of brilliant minds at one institution"* (Kimball, 1997). Beyond just impacting their experience on the NIH campus, the density of talent at the NIH helped form a rich network alumni could tap into later in their careers.

Finally, many ATP fellows came to view the focus on what would later be called translational research as a distinctive element of the approach to research at the NIH. This was no accident. James Shannon, one of the early leaders of the NIH, carefully structured the intramural program to facilitate close cooperation between basic and clinical research (Goldstein and Brown, 1997; Park, 2003). This was reinforced in the physical design of buildings, where the clinical center had patient care areas and research laboratories adjacent to one another within each floor to facilitate interactions across them. Shannon saw a special role for the physician-scientist, who could make fundamental discoveries about biologic mechanisms and apply these findings to the bedside. Anthony Fauci, an ATP alumni and prominent HIV/AIDS researcher, recalls, *"what the Clinical Associate Program does is it gives you a very interesting perspective on the relationship between disease and the basic science that you have to study to be able to approach disease... Also the link, as we used to say, between 'the bed and the bench,' you see something at the bedside, you bring it back and ask the question at the bench or you make a discovery at the bench and you go back and apply it to the bedside, that bedside to bench phenomena was really what the Clinical Associates program was all about"* (Fauci, 1998). Fauci also contrasts the approach at the NIH to those of other academic medical centers at the time, *"[At Cornell] there was very little time to think about why patients developed certain diseases or infections. It was always treat them, get them ready and get them out. Whereas at the NIH, you see the patient and then you say, 'You know, I think I want to do a project to ask that question.'"*

The Doctor Draft. Understanding the institutional setting of the NIH ATP requires reviewing the relationship between physicians and the military draft during the Korean and Vietnam Wars. With the start of the Korean War in June of 1950, there was an increased need for physician in the military to care for the expanding population of enlisted personnel. While physicians could be drafted, their extended training and educational exemptions in the draft allowed them a better chance to avoid being called through the standard draft process (Card and Lemieux, 2001; Park, 2003). To ensure an adequate supply of physician draftees, lawmakers passed an amendment to the Selective Service Act in September of 1950, colloquially known as the Doctor Draft. This amendment allowed for the special registration and induction of physician and certain allied health professionals. The Berry Plan, named after Assistant Secretary of Defense for Health and Medical Affairs Frank Berry, was enacted in 1954 and modified the Doctor Draft. The Berry Plan allowed trainees to defer their military service during medical school and for a certain portion of their residency training; however, timing preferences were often not honored (Berry, 1976; Klein, 1998). In 1967 there were additional restrictions on exemptions for physicians seeking draft deferment. This change in exemptions, along with increasing needs by the Department of Defense, led to about 6,000 of the 9,000 doctors graduating from medical school annually to be drafted, with about 700 physicians, dentists, and allied professionals able to fulfill military obligations through the Public Health Service Commissioned Corps (Committee on Labor and Public Welfare, 1970).

Not only did the draft influence individual decisions to apply to the NIH ATP, it shaped the growth of the NIH by facilitating concentration of talent in one location. Some have argued that the ATP and doctor draft played an important role in the preeminence of the NIH for precisely this reason (Park 2003). Harry Kimball, an alumni of the program and former president of the American Board of Internal Medicine said, *"We all knew we were going to serve in the military one way or the other... so it was just a matter of trying to arrange the best possible experience during your military time... the fact that there was a doctor draft made the NIH the premier place."* (Kimball, 1997).

Program cost and return on investment. Unfortunately, only limited information is available to us to estimate the cost to the NIH of running the ATP: the NIH does not disclose many of the details regarding the internal structure of its intramural budget. Historically, the relative size of the intramural program has waxed and waned between 10 and 15% of the overall NIH budget. The intramural program was 9.6% of the NIH budget in 1977, 10.1% in 1978, 10.8% in 1979, and 11.1% in 1980 (Institute of Medicine, 1988); comparable figures for earlier years are not available to us.

While the actual cost of the ATP program is not known, we can make a back of the envelop calculation of the costs. To begin with, one might start with the fully-loaded cost of a postdoctoral fellow in 2019, which we estimate at $100,000 per year as a ballpark figure.[iv] Using the size of each cohort of trainees (including trainees not part of our sample as discussed in section 3.1) and the Biomedical Research Price Index, it is straightforward to generate a rough estimate of the program budget. We find the overall cost of the program to fall within a range of 2.4 to 5.6 million current dollars (or 27 to 44 million constant 2019 dollars). In turn, this represents between 2 and 5% of the intramural budget, and between 0.25% and 0.4% of the overall NIH budget during the years of our study.

These costs strike us as small, akin to a drop in the proverbial bucket. Of course, working out a rate of return on these training investments requires many additional—and untestable—assumptions, in particular whether the inventions linked to the discoveries of ATP trainees would have happened at all, or happened later in the absence of the program. As usual, in a portfolio of projects whose returns are extremely skewed, even under very conservative assumption, one key invention is all that is needed to justify the cost of the program many times over. In our particular case, the invention of statin drugs, which follows quite directly from the research of ATP fellows Joseph Goldstein and Michael Brown, provides just the example needed to argue that the social rate of return on NIH training investments during this period is likely to be enormous.

References

Card, David, and Thomas Lemieux. 2001. "Going to College to Avoid the Draft: The Unintended Legacy of the Vietnam War." *American Economic Review* **91**(2): 97-102.

Cassell, Gail H. et al. 1994. "NIH Intramural Research Program. Report of the External Advisory Committee of the Director's Advisory Committee."

Committee on Labor and Public Welfare. Subcommittee on Health. 1970. "National Health Service Corps Act of 1970." Washington, DC: U.S. Government Printing Office.

DeVita, Jr., Vincent T. 1997. "National Cancer Institute Oral History Project." edited by Gretchen A. Case. Office of NIH History, National Institutes of Health.

[iv] There are reasons for a year of fellowship at NIH to be more costly than this, perhaps because of the fixed costs associated with running the program, or because NIH trainees need more "hand-holding" early on given their lack of research experience. On the other hand, we could also see the case for the number to be lower, since all trainees are physicians, and the clinical associates (a third to a half of the trainees) perform clinical duties within the patient floors of the NIH clinical research centers, and these services have a direct market value for the institution.

Gellhorn, Alfred. 1965. “Research at NIH: The Wooldridge Report.” *Science* **149**(3679): 6.

Institute of Medicine. 1988. “Healthy NIH Intramural Program, Structural Change or Administrative Remedies: Report of a Study by a Committee of the Institute of Medicine, Division of Health Sciences Policy.” Washington, DC: The National Academies Press.

NIH Office of History. 1982. “Intramural Science at the NIH.” Office of NIH History, National Institutes of Health.

Schechter, Alan, and Geraldine Schechter. 1998. “Clinical Associates Program Oral History Series.” edited by Melissa Klein. Office of NIH History, National Institutes of Health.

Scolnick, Edward. 1998. “National Cancer Institute Oral History Project.” edited by Gretchen A. Case. Office of NIH History, National Institutes of Health.

Shapiro, Harold T. et al. 1988. “A Healthy NIH Intramural Program: Structural Change or Administrative Remedies? Report of a Study.” Institute of Medicine.

Wooldridge Committee Report. 1965. “Biomedical Science and Its Administration: A Study of the National Institutes of Health.” edited by President’s NIH Study Committee. Report to the President. February 1965 United States. Government Printing Office.

Appendix F
Econometric Considerations

Inverse probability of treatment weighted estimation. Let us first assume that the NIH principal investigators recruiting fellows at the interview stage are unable to select applicants on the basis of covariates unobserved by the econometrician and correlated with research career success—the "unconfoundedness" assumption. This assumption is not refutable and it places strong demands on the data generating process.[v]

In addition, we must assume that, for all included values of the covariates predicting treatment, the likelihood of being selected to attend is positive—the "common support" assumption. The common support assumption implies that we should limit our comparisons to sets of values for which there is sufficient overlap in the match probabilities between actual and counterfactual matches (Barber et al., 2004).

Under these assumptions, Hirano and Imbens (2001) show that various treatment effects of attending the NIH ATP on outcome y, conditional on exogenous applicant characteristics Z, can be recovered by estimating

$$E[y|X,Z] = \beta_0 + \beta_1^{'}Z + \beta_2 TREAT \tag{E.1}$$

by weighted least squares or weighted maximum likelihood (depending on the distribution of y), where the weights correspond to the inverse probability that each observation is treated. Implementation is straightforward. We first estimate the propensity of attending the program as a function of pre-treatment observable characteristics

$$\phi(X_i) = Prob(TREAT_i = 1|X_i) \tag{E.2}$$

for each applicant i. The predicted probabilities (the propensity scores) help create regression weight w_i for each subject. To estimate the Average Treatment Effect (ATE):

$$w_i = \begin{cases} \frac{1}{1-\hat{\phi}(X_i)} & \text{if} \quad TREAT_i = 0 \\ \frac{1}{\hat{\phi}(X_i)} & \text{if} \quad TREAT_i = 1 \end{cases} \tag{E.3.1}$$

To estimate the Average Treatment Effect on the Treated (ATET):

$$w_i = \begin{cases} \frac{\hat{\phi}(X_i)}{1-\hat{\phi}(X_i)} & \text{if} \quad TREAT_i = 0 \\ 1 & \text{if} \quad TREAT_i = 1 \end{cases} \tag{E.3.2}$$

Weighting equation (E.1) by w_i effectively creates a pseudo-population of applicants in which X no longer predicts assignment to treatment and the causal association between treatment and the outcome variable is unchanged from the original population.[vi] We refer to β_2 when equation (E.1) is weighted by w_i as the Inverse Probability of Treatment Weighted (IPTW) estimator of β_2 (Austin and Stuart, 2015; Xu et al., 2010).

Informative censoring. Although we focused the first part of the discussion on the problem of non-random selection into treatment, a second problem arises because some applicants might fail to engage in research

[v] We know from past research that "selection-on-observables"-type techniques perform best (in the sense of replicating an experimental benchmark) when it is possible to include a comprehensive list of covariates to model the probability of assignment to treatment (Dehejia and Wahba, 2002). In our sample, we have at our disposal a large set of pre-treatment covariates that we believe to be likely to confound comparisons between attendees and non-attendees: quality of medical school attended, research publications as an undergraduate and medical school student, etc.

[vi] We use "stabilized" inverse probability of treatment weighting, which multiplies weights by the marginal probability of receiving treatment. This method addresses the difficulty of very large weights being assigned to treated individuals with a probability of treatment close to zero or controls with a probability of treatment close to one (Austin and Stuart, 2015; Xu et al., 2010).

activities for the sole reason that their chosen position does not afford them the possibility to publish, seek external grants, or train the next generation of scientists. This problem is distinct from informative loss to follow-up. These physicians' careers are observed in full and yet it does not seem meaningful to compare the research productivity of a full-time, tenure-track academic researcher with that of a clinician who very occasionally dabbles in research. We deal with this problem by treating early exit from research as another treatment. As Robins et al. (2000) note, adjusting for this type of informative censoring in this way is tantamount to estimating the causal effect of ATP attendance on an outcome if, contrary to the fact, all applicants had remained engaged in research rather than followed their censoring history. We model the exit decision as a function of the same pre-application covariates used to model selection into treatment, and compute weights corresponding to the probability of exit given these observables. Concretely, we estimate the propensity of early exit as a function of pre-treatment observable characteristics and program receipt:

$$\rho(X_i) = Prob(EXIT_i = 1|X_i) \tag{E.4}$$

for each applicant i. The predicted probabilities help create regression weights v_i for each subject. For example, the Average Treatment Effect (ATE) can be estimated with v_i defined as:

$$v_i = \begin{cases} \frac{1}{1-\hat{\rho}(X_i)} & \text{if} \quad EXIT_i = 0 \\ \frac{1}{\hat{\rho}(X_i)} & \text{if} \quad EXIT_i = 1 \end{cases} \tag{E.4.1}$$

Hernan et al. (2001) show that consistent estimates for β_2 can be obtained by multiplying the weight corresponding to the inverse probability of treatment w_i and the weight corresponding to the inverse probability of censoring v_i. The denominator of the final weight is the probability that an applicant subject would have followed his own treatment and censoring history, conditional on observables. We label this methodology Inverse Probability of Treatment and Censoring-Weighted (IPTCW) estimation in what follows.

References

Barber, Jennifer S., Susan A. Murphy, and Natalya Verbitsky. 2004. "Adjusting for Time-Varying Confounding in Survival Analysis." *Sociological Methodology* **34**(1): 163-192.

Dehejia, Rajeev H. and Sadek Wahba. 2002. "Propensity Score-Matching Methods for Nonexperimental Causal Studies." *The Review of Economics and Statistics* **84**(1): 151-161.

Hernan, Miguel A., Babette Brumback, and James M. Robins. 2001. "Marginal Structural Models to Estimate the Joint Causal Effect of Nonrandomized Treatments." *Journal of the American Statistical Association* **96**(454):440-448.

Appendix G
Research Style Measures

To characterize research style, we take advantage of MeSH terms, a hierarchical controlled vocabulary thesaurus maintained by the National Library of Medicine. The National Library of Medicine employs professional indexers to select MeSH indexing terms for biomedical publications according to specific protocol and considers each article in the context of the entire collection. Importantly, given the subjectivity of any indexing task, the authors are not involved in the process of selecting MeSH terms.

Disease-oriented articles were identified as all those under the MeSH tree disease category (C01-C26). We excluded C22, which primarily measures veterinary diseases, and included F03, mental disorders. All together this measure includes 4,895 unique MeSH terms. Example terms include *hematuria*, *aortic valve stenosis*, and *Klippel-Feil Syndrome*. Some microbiologic agents, such as *escherichia coli*, may be both pathologic causes of disease as well as common organisms in molecular biology research. To ensure results are not being driven by conflation of microbiologic disease with a research model organism, a second measure of disease-orientation was constructed as above but dropping all bacterial infectious disease terms (C01) with similar results obtained.

Articles using molecular biology methods were identified primarily based on the MeSH category for investigative techniques (F05). These were manually reviewed to eliminate any terms which may have a potential direct clinical application outside of a laboratory, such as *angioplasty* or *glasgow coma scale*. The final list of MeSH codes includes: E05.017, E05.091, E05.118, E05.181, E05.196 (but excluding E05.196.353), E05.197, E05.198, E05.242.223, E05.242.335, E05.242.363.342, E05.242.373, E05.242.383.910, E05.242.551, E05.242.800, E05.295, E05.301, E05.313, E05.318.416, E05.393, E05.478, E05.481, E05.484, E05.490, E05.522, E05.588, E05.591, E05.595, E05.598, E05.601, E05.624, E05.650, E05.657, E05.830, E05.916.680, and A11.251 (excluding A11.251.476). Together, these codes identify 510 unique MeSH terms with examples including *immunoelectrophoresis*, *nucleic acid hybridization*, and *real-time polymerase chain reaction*.

To construct a measure of the use of a model organism in research, we compiled a list of 53 different model organisms used in biomedical research: *tobacco mosaic virus*, *bacteriophage* λ, *bacteriophage* $\phi X174$, *SV40*, *T4 phage*, *escherichia coli*, *bacillus subtilis*, *caulobacter crescentus*, *aliivibrio fischeri*, *synechocystis*, *pseudomonas fluorescens*, *azotobacter vinelandii*, *Streptomyces coelicolor*, *chlamydomonas reinhardtii*, *dictyostelium discoideum*, *tetrahymena thermophila*, *eremothecium*, *aspergillus nidulans*, *coprinus cinereus*, *Cryptococcus neoformans*, *ceurospora crassa*, *saccharomyces cerevisiae*, *schizophyllum*, *schizosaccharomyces pombe*, *ustilago maydis*, *arbacia punctulata*, *aplysia*, *caenorhabditis elegans*, *ciona intestinalis*, *drosophila*, *loligo pealei*, *trichoplax adhaerens*, *ambystoma mexicanum*, cat, chicken, dog, *mesocricetus auratus*, guinea pigs, rabbits, *oryzias latipes*, mice, genetically modified animals (B01.050.050.136 and B01.050.050.199.520), mole-rat, pigeon, *poecilia reticulata*, rat, *rhesus macaque*, *petromyzon marinus*, *takifugu*, *xenopus laevis*, and zebrafish. The associated MeSH codes for these organisms resulted in 125 unique MeSH terms. As a robustness check to ensure the result was not driven by microbiologic model organisms which may also be pathologic diseases, as well as for clarity given the large number of organisms, a second measure limited only to major non-microbiologic organisms was constructed. This consisted of the MeSH codes for the following model organisms: *caenorhabditis elegans*, *drosophila*, zebrafish, mice, genetically modified animals (examples under this category include knockout and SCID mice), *saccharomyces cerevisiae*, and *rhesus macaque*. This subset contained a total of 67 unique MeSH terms. Results were similar using this alternative measure of model organism.

We constructed two additional measures of basic science based on the research topic: the first focused on cellular structures and macromolecules, and the second on biochemical and cellular processes. We identified 2,620 MeSH terms related to cellular structures and macromolecules and 1,028 related to biochemical and cellular processes. Care was taken to avoid terms which may have direct clinical relevance. The final list of MeSH codes for cellular structures and macromolecules includes: A11.284, A20 (excluding A20.593), D05.500, D08.811, D09.067 (excluding D09.067.687.668, D09.067.342.531), D09.254, D12.125.780, D12.644.360, D12.644.770, D12.776.575, D12.776.580, D12.776.835, D12.776.938, D12.776.947, D13, D23.125, and D23.585. Example MeSH terms capture by this measure include *golgi apparatus*, *16S ribosomal RNA*, *DNA topoisomerase IV*, and *COP9 signalosome complex*. The final list of MeSH codes for biochemical and cellular processes includes: G02 (but excluding G02.111.130, G02.111.007, G02.186, G02.819), G03 (but excluding G03.015, G03.030, G03.180, G03.191, G03.312, G03.442, G03.458, G03.615.500, G03.680, G03.787, G03.800, G03.820, G03.857), G04 (excluding G04.140), G05 (excluding G05.045, G05.090, G05.180, G05.285, G05.347, G05.350, G05.390, G05.400, G05.410, G05.697, G05.815, G05.910), G06.920 (excluding G06.225.420, G06.920.850), G07.265.755, G12, G11.561.653, G11.561.638, G16.075.250. Examples include *chaperone-mediated autophagy*, *signal transduction*, *post-transcriptional RNA processing*, and *oxidative phosphorylation*.

Clinical trial articles were identified by two approaches. First, we used MeSH terms for publication type (V03.175) as well as topic (E05.318.372.250, N05.715.360.330.250, N06.850.520.450.250), with veterinary terms eliminated (V03.175.375, V03.175.750, N05.715.360.330.250.375, N06.850.520.450.250.750). These MeSH codes resulted in 25 unique MeSH terms. Examples include *clinical trial, phase II*; *randomized controlled trial*; and *observational study*. As a second measure, we identified all those papers tagged in *PubMed* with a publication type containing the term trial. Example publication types include *adaptive clinical trial*; *clinical trial, phase III*; and *randomized controlled trial*.

Appendix H
Identification of Trainees

To identify the trainees of ATP applicants who later go on to receive NIH grants, we first identified the set of original research articles after career independence in which the ATP applicant was the last author. The first authors of these publications were then matched against NIH grantees who earned their doctorate degree 1965-2015 from the Consolidated Grant Applicant File and NIH Exporter. Only those publications that occurred in a window centered on the time of his or her highest degree were considered to be during training (3 years prior to earning a doctorate to 5 or 7 years afterwards for a PhD or MD, respectively, to account for residency, fellowship, and postdoctoral training as well as any publication lags). This established a set of potential trainee/ATP applicant dyads.

We employed several strategies to verify the potential match between the first author and NIH grantee. We considered all dyads matching on location or specialty to be a valid match. We defined a location match as occurring if the NIH grantee institution matched either the institutions of the NIH ATP applicant's first or last job after career independence. This approach capitalizes on the fact that a significant number of residency or fellowship graduates take their first faculty job at the same institution they completed their training. Care was taken to account for close institutional affiliations (for example, the San Francisco Veterans Affairs Medical Center and University of California, San Francisco would be considered as matching).

The specialty of NIH grantees was derived from their departmental affiliation when available. A difficulty with this approach is that some medicine subspecialists are reported as working in the general medicine department rather than their subspecialty (i.e. reported in the department of medicine rather than a cardiology department). We conservatively only considered exact specialty matches (i.e. general internal medicine matching to general internal medicine and cardiology to cardiology). As a robustness check, the analysis was repeated using a stricter definition of specialty match which excluded any general medicine or general pediatrics matches. We also considered a specialty as matching if a pediatric subspecialist studied with their adult physician counterparts or vice versa.

For our baseline specification, we defined a trainee/ATP applicant dyad as valid if it matched on location, matched on specialty, or had a last name frequency within the NIH grantees of less than or equal to 10. As a robustness check, we employed a stricter definition using only a hand-coded subset. We consider R01 grantees to also include those receiving R29 and R37 grants.