

Essays on Biomedical Innovation

By

Wesley H. Greenblatt

B.S., Yale University, 2007

M.S., Yale University, 2007

M.D., Harvard Medical School, 2012

S. M., Massachusetts Institute of Technology, 2021

SUBMITTED TO THE DEPARTMENT OF MANAGEMENT IN PARTIAL
FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN MANAGEMENT

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2023

©2023 Wesley H. Greenblatt. All rights reserved.

The author hereby grants to MIT a nonexclusive, worldwide, irrevocable, royalty-free license to exercise any and all rights under copyright, including to reproduce, preserve, distribute and publicly display copies of the thesis, or release the thesis under an open-access license.

Authored by: Wesley H. Greenblatt
MIT Sloan School of Management
April 11, 2023

Certified by: Pierre Azoulay
International Programs Professor of Management, Thesis Supervisor

Certified by: Scott Stern
David Sarnoff Professor of Management of Technology, Thesis Supervisor

Accepted by: Eric So
Sloan Distinguished Professor of Financial Economics
Professor, Accounting and Finance
Faculty Chair, MIT Sloan PhD Program

Essays on Biomedical Innovation

by

Wesley H. Greenblatt

Submitted to the Department of Management on April 11, 2023, in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy in Management

ABSTRACT

The choice of what scientists decide to study is fundamental to determining what innovations are produced. Yet, rather than being solely determined by a scientist's idiosyncratic preferences, this choice is shaped in important ways by the incentives, organization environment, and supporting institutions that surround a researcher. This dissertation consists of three essays that use biomedicine as a setting to explore how institutions and organizational environment can shape the rate and direction of innovation.

In the first essay, I examine how knowledge certification by professional medical society clinical practice guidelines shapes the use of extant knowledge. Employing a difference-in-differences approach, I find that after inclusion in a guideline affected subfields of knowledge grow in size and scientific impact compared to controls carefully matched on observables. Rather than the aperture of subsequent innovation narrowing, subfields shift towards exploration as they become more translational, more intellectually distant, more disruptive, and build on more diverse and less established prior research.

In the second essay, I investigate how exposure to frontier research in a training program can alter the career trajectories of potential innovators. I study the careers and innovative output of physicians who applied to the Associate Training Programs of the National Institutes of Health (NIH) during the turbulent period surrounding the Vietnam War. I find program participants entered research-focused positions at higher rates and garnered more publications, citations and grant funding than synthetic controls. In particular, the direction of their research efforts was durably imprinted with a distinct "translational" style of biomedical research that was characteristic of the NIH at the time.

In the third essay, I study how a specific institution, grant peer review, shapes scientific risk taking. While risk is an inherent aspect of innovation, those projects with high degrees of risk may be more likely to lead to breakthrough innovation yet may face challenges in winning the support necessary to be carried out. I analyze R01-equivalent grants from the NIH, and after carefully controlling for investigator, grant and institution characteristics, find that grants with high levels of risk taking are renewed at lower rates.

Thesis Supervisor: Pierre Azoulay (Co-chair)

Title: International Programs Professor of Management

Thesis Supervisor: Scott Stern (Co-chair)

Title: David Sarnoff Professor of Management of Technology

Acknowledgements

I am immensely grateful to have had the opportunity to undertake doctoral studies in the TIES community. This is a path I couldn't have taken without the support of many individuals, who not only helped me to grow as a scholar, but also made it such a rich, rewarding, and fun journey along the way.

First, I am deeply grateful to my committee members: Pierre Azoulay, Scott Stern, and Fiona Murray. Words hardly reflect the extent of my gratitude. Each were generous with their time, and always seemed to strike the perfect balance between challenging me to improve my work while being supportive through the ups and downs of research. More than that, I learned an enormous amount by the examples each set as a scholar, educator, and colleague.

Pierre has been instrumental in my development. He helped teach me rigorous empirical methods and the attention to detail needed to produce high quality research. In particular, Pierre was patient with me and gave detailed feedback as I learned Stata, management scholarly writing, and new empirical methods, especially considering my limited economics background when I entered the program. But most of all I am grateful for how dedicated he has been as a mentor: whether it was the evening zooms while on vacation with his family or trips into Sloan on the weekend to meet with me, Pierre has truly gone above and beyond in supporting me.

Scott urged me to think about the big picture of my research and which questions I should be asking to have an impact. In the process, he encouraged me, with his infectious enthusiasm, to leave my computer and talk to practitioners to make sure I really understood the phenomenon and use this to guide my research. Without fail, I left his office with a deeper understanding of my research and new questions to pursue.

Fiona had an unparalleled ability to immediately cut through complex problems to the core issue at hand. She would frequently offer the perspective I needed to quickly solve problems I had previously spent days trying to address. She also was extremely supportive, especially during the pandemic, in helping me remain grounded and connected while at Sloan.

I also want to thank my coauthors during my doctoral studies: I have learned so much from working on papers together. In addition to Pierre, I have been fortunate to collaborate with Jennifer Kao, Charu Gupta, Roger Levy, Suman Maity, and Misty Heggeness. Misty in particular was especially understanding of balancing the demands of my doctoral studies with our work at the U.S. Census Bureau and was always guided by what was best for my professional development when conflicting demands emerged. I was also fortunate to have the wonderful research assistantship from Yuxin Chen and Aaron Robles.

I couldn't have completed the Ph.D. without the help and support of the other TIES students. They helped solve questions large and small without which progress would have been much slower. But more importantly, they helped create a warm, supportive community which made me excited to come to Sloan each day. This includes Ankur Chavda, Alexis D'Alessandro, Caroline Fry, Luca Guis, Jorge Guzman, Mike Kearney, Danny Kim, Hyejun Kim, Soomi Kim, Josh Krieger, Christina Nguyen, Lindsey Raymond, Laura Shupp, Jane Wu, and Sam Zyontz.

I am also grateful to the wider community at TIES and MIT. I was fortunate to receive feedback and guidance on research and career development from Christian Catalini, Mercedes Delgado, Danielle Li, Jacquelyn Pless, Georg Rilinger, Ben Roin, Neil Thompson, Jim Utterback, and Eric von Hippel. I have also had the privilege of learning pedagogy from Bill Aulet, Pierre Azoulay, Trish Cotter, Kit Hickey, Michellana Jester, Simon Johnson, Hiram Samel, Scott Stern, and especially Erin Scott. Judy Graham-Robey, Natalia Kalas, Natalie Jean, Kim McGrath and Danielle Tarpley have been instrumental in helping me navigate both MIT and the job market. I'm also grateful to Hillary Ross and Davin Schnappauf for all they do to support the Ph.D. program and always being an advocate for the doctoral students.

While MIT was the start of studying innovation and entrepreneurship, my prior research mentors played an important role in preparing me for doctoral studies and instilling in me a passion for research. This includes Bala Chandran, Scott Gazelle, Theodore Marmor and especially Mark Hochstrasser and Chin Hur. I'm also grateful to Jeff Furman for encouraging me to apply to the Ph.D. program despite my non-traditional career path as well as Richard Bachur and Jason Levy at Boston Children's Hospital for their support of my career vision.

Finally, no thanks would be complete without acknowledging my family. In particular, none of this would have been possible without the support of my parents. My brother Matt, a physician-scientist, gave invaluable perspective on navigating doctoral studies as a physician. Most importantly, I am grateful to my wife Louise for undertaking this journey together with me.

Table of Contents

Chapter 1. Introduction	11
References	14
Chapter 2. Anchoring Exploration: Knowledge Certification and the Dynamics of Innovation.....	15
1. Introduction	16
2. Knowledge certification and innovation.....	19
3. Clinical practice guidelines	23
History and evolution.....	24
CPGs and the study of innovation.....	25
4. Empirical design, data and descriptive statistics.....	26
Data and measurement.....	26
Empirical strategy	30
Econometric considerations.....	31
Descriptive statistics.....	32
5. Results	33
Effect of CPG on subfield growth, attention, and scientific impact.....	33
CPGs and the direction of research within a subfield.....	34
Commercial orientation of subfield research.....	36
Which subfields are most impacted by CPGs?	36
Alternative mechanisms.....	37
Robustness analysis	38
6. Discussion	40
References	45
Figures.....	51
Tables.....	53
Appendix A. Ancillary Results and Robustness Checks.....	61
Appendix B. Clinical Practice Guidelines.....	78
Appendix C. Measurement and Methods.....	80
Chapter 3. Long-term Effects from Early Exposure to Research: Evidence from the NIH “Yellow Berets”	88
1. Introduction	89
2. Institutional setting	92
The Associate Training Program.....	92
The application process	94
Prior evaluations.....	95
3. Empirical design, data and descriptive statistics.....	96
Data.....	96
Descriptive statistics.....	97
Econometric considerations.....	102
4. Results	103
Selection into the NIH ATP.....	104
Career choice	105
Research outcomes.....	106
Research style.....	108
Mechanisms	108
Robustness analyses.....	111
5. Conclusion.....	112
References	116
Figures.....	120

Tables.....	122
Appendix A. Data Set Construction.....	132
Appendix B. Robustness Checks and Ancillary Results.....	133
Appendix C. Non-applicant Sample.....	157
Appendix D. Draft Lottery Subsample.....	159
Appendix E. Background Institutional Detail.....	161
Appendix F. Econometric Considerations.....	165
Appendix G. Research Style Measures.....	167
Appendix H. Identification of Trainees.....	169
Chapter 4. Does Grant Peer Review Penalize Scientific Risk Taking? Evidence from the NIH.....	170
1. Introduction.....	171
2. Peer review and risk taking.....	173
3. Empirical design, data and descriptive statistics.....	176
Setting: The National Institutes of Health.....	176
Data.....	177
Measuring risk taking.....	178
Empirical strategy.....	181
Econometric considerations.....	182
Descriptive statistics.....	183
4. Results.....	184
Risk taking and grant renewal.....	184
Novelty and risk taking.....	184
Who is penalized for risk taking?.....	185
Robustness analysis and ancillary results.....	185
5. Discussion.....	187
References.....	191
Figures.....	195
Tables.....	203
Appendix A. Ancillary Results and Robustness Checks.....	211

List of Figures

Chapter 2

Figure 1. Distribution of pre-treatment root article characteristics.....	51
Figure 2. Dynamic effects of CPG on subfield growth	52
Figure 3. Effect of CPG on subfield scientific impact.....	52
Figure A1. Clinical practice guideline characteristics over time.....	63
Figure A2. Raw annual subfield article count over time	64
Figure A3. Distribution of subfield overlap	64
Figure A4. Effect of CPG on root article citations.....	65
Figure A5. Field trends over time: publication counts	66
Figure A6. Field trends over time: patent counts.....	67
Figure C1. Subfield construction and evolution.....	85
Figure C2. Illustration of treated and control root articles	85

Chapter 3

Figure 1. NIH ATP interviewed candidates by year	120
Figure 2. Predicted probability of selection	121
Figure B1. Pre-application publications.....	137
Figure B2. Career research outcomes.....	138
Figure B3. Predicted probability of selection for logit specification	139
Figure B4. Specialization and career research productivity	140
Figure B5. Treatment effect dynamics.....	141
Figure C1. Age at medical school graduation	158

Chapter 4

Figure 1. Distribution of grant cycle publications by renewal status	195
Figure 2. Distribution of risk-taking measures by renewal status	196
Figure 3. Effect of risk taking on grant renewal: Extreme tail outcomes	197
Figure 4. Effect of risk taking on grant renewal: Disruption	198
Figure 5. Effect of risk taking on grant renewal: Pivoting	199
Figure 6. Effect of risk taking on grant renewal: Standing out	200
Figure 7. Effect of risk taking and novelty on grant renewal	201
Figure 8. Heterogeneity in penalty for risk taking.....	202
Figure A1. Distribution of measures of novelty	211
Figure A2. Effect of novelty on grant renewal.....	212
Figure A3. Distribution of alternative measures of extreme tail outcomes	213
Figure A4. Effect of risk taking on grant renewal: Alternative measures of extreme tail outcomes	214
Figure A5. Distribution of alternative measures of disruption	216
Figure A6. Effect of risk taking on grant renewal: Alternative measures of disruption	217
Figure A7. Distribution of alternative measures of pivoting	218
Figure A8. Effect of risk taking on grant renewal: Alternative measures of pivoting	219
Figure A9. Distribution of alternative measures of standing out	220
Figure A10. Effect of risk taking on grant renewal: Alternative measures of standing out.....	221

List of Tables

Chapter 2

Table 1a. Descriptive statistics: Guidelines	53
Table 1b. Descriptive statistics: Guideline series.....	53
Table 2. Pre-treatment characteristics of treated and control root articles.....	54
Table 3. Effect of CPG on subfield growth and citations.....	55
Table 4. Effect of CPG on subfield research direction.....	56
Table 5. Effect of CPG on subfield commercial orientation	57
Table 6. Effect of CPG by subfield maturity.....	58
Table 7a. Effect of CPG by last author expertise within cardiology.....	59
Table 7b. Effect of CPG by root article prominence.....	59
Table 8. Effect of CPG by previous codification in review articles.....	60
Table A1. Descriptive statistics: Guidelines compared to review articles.....	68
Table A2. Covariate balance across different treatment and control samples.....	69
Table A3. Alternative measures of research direction	70
Table A4a. Effect of CPG by last author expertise.....	71
Table A4b. Effect of CPG by within subfield authors.....	71
Table A5. Dose effect.....	72
Table A6. Placebo tests	72
Table A7a. Effect of CPG inclusion on root article citations	73
Table A7b. Effect of CPG inclusion on author and journal diversity	73
Table A8. Alternative control groups	74
Table A9. Exclusion of articles citing the CPG or root article.....	75
Table A10. Alternative functional forms	75
Table A11. Two-period OLS.....	76
Table A12. Analysis at level of CPG.....	76
Table A13. Stacked regression	77
Table C1. List of sponsoring organizations.....	86
Table C2. Distribution of cardiovascular disease categories across samples.....	87

Chapter 3

Table 1a. Descriptive statistics: Pre-application data	122
Table 1b. Descriptive statistics: Career choice	123
Table 1c. Descriptive statistics: Research outcomes.....	123
Table 1d. Descriptive statistics: Notable achievements.....	124
Table 1e. Descriptive statistics: Research style	124
Table 2. Modeling selection into the NIH ATP.....	125
Table 3. Career choice outcomes	127
Table 4. Research outcomes.....	128
Table 5. Publication outcomes, by citation quantiles.....	129
Table 6. Research style.....	130
Table 7. Robustness analysis	131
Table B1. Most common medical schools attended by applicants.....	142
Table B2. Occupational breakdown, first position (post-training)	143
Table B3. Occupational breakdown, last position	144
Table B4a. Covariate balance using logit weights.....	145
Table B4b. Research and training outcomes using logit weights.....	146
Table B5. Exclusion of controls not receiving research training.....	147
Table B6. Intensity of treatment effects (dose-response relationship).....	148
Table B7a. Descriptive statistics: Mentor outperformance.....	149
Table B7b. Mentor outperformance.....	149
Table B8a. Descriptive statistics: Coauthorship peer effects.....	150

Table B8b. Coauthorship peer effects.....	150
Table B9. Researcher specialization.....	151
Table B10. Research and training outcomes by weighting procedure	152
Table B11. Research outcomes and style during training, by program track.....	153
Table B12. Treatment effect by application year.....	154
Table B13. Research outcomes and style, by program track.....	155
Table C1. Descriptive statistics: Characteristics of non-applicants.....	158
Table D1. Modeling selection into the NIH ATP: Draft Lottery	160

Chapter 4

Table 1a. Descriptive statistics: Investigator characteristics.....	203
Table 1b. Descriptive statistics: Grant cycle	204
Table 1c. Descriptive statistics: Measures of risk taking and novelty	205
Table 1d. Descriptive statistics: Correlations across measures of risk taking and novelty.....	205
Table 2. Effect of risk taking on grant renewal	206
Table 3. Effect of novelty on the risk taking grant renewal penalty	210
Table A1. Effect of novelty on the risk taking grant renewal penalty	222
Table A2. Descriptive statistics: Alternative measure of risk taking and novelty.....	225
Table A3. Heterogeneity by degree of pivoting in proposed grant abstract	226
Table A4. Robustness: Investigator name frequency	227
Table A5. Robustness: Reliance on NIH funding.....	228
Table A6. Robustness: Time period.....	229
Table A7. Robustness: Oster’s delta.....	230

Chapter 1

Introduction

The choice of what scientists decide to study is fundamental to determining what innovations are produced. Yet, rather than being solely determined by a scientist’s idiosyncratic preferences or brilliance, this choice is shaped in important ways by the incentives, organization environment, and supporting institutions that surround a researcher.

A growing literature has explored how such factors shape not just the amount of innovation, but also its direction. Examples include studies of the structure of financial incentives (Budish et al. 2015, Graff Zivin and Lyons 2020), firm financial resources (Krieger et al. 2021), tolerance for failure (Azoulay et al. 2011, Tian and Wang 2014), access to data and research samples (Furman and Stern 2011, Murray et al. 2016, Nagaraj et al. 2020), technological tools for discovery (Furman and Teodoridis 2020, Kim 2022), team structure (Lee et al. 2015, Wu et al. 2019), and colleague spatial proximity (Catalini 2018, Lee 2019).

This dissertation contributes to this literature with three essays that use biomedicine as a setting to explore how institutions and organizational environment can shape the rate and direction of innovation. A key lesson is that emerges from this work is that the desired impact on innovation should be built into the design of supporting institutions and organizational environments.

Biomedicine is an attractive setting for studying these issues for a number of reason. First, it has rich data available. Scientific norms lead to most innovations being published, and patents play a prominent role in biomedicine in protecting innovations. Additionally, in many sectors within biomedicine, there is also extensive data available on the careers of innovators, including professional activities outside of research, as well as clearly defined populations of those “at risk” of innovating. Second, biomedicine presents a fertile environment for studying these issues, as innovation may occur under a diverse set of institutional and organizational regimes. There is a large role for both academic and industry innovation, facilitating study along the entire spectrum from idea production to successful commercialization. Additionally, there are many prominent professional and governmental institutions and policies which may impact innovation. Together, these two factors suggest biomedicine is an appealing setting to study due to the empirical window it can provide on topics of broad interest outside of biomedicine in the strategic management of innovation. Finally, as healthcare is substantial share of the overall economy and biomedical innovations may improve the quality and length of life, a deeper understanding of biomedical innovation is an important topic to study in and of itself.

In the first essay, “Anchoring Exploration: Knowledge Certification and the Dynamics of Innovation,” I study how inclusion of extant knowledge in a professional medical society clinical practice guideline—a form of knowledge certification—shapes follow-on innovation. Rather than focusing on citations to trace follow-on innovation, I delineate subfields of research articles based on shared intellectual content and characterize how these subfields evolve over time. Employing a difference-in-differences estimator, I studying the within-subfield change in research following inclusion of an article in a clinical practice guideline compared to articles and their associated subfields, carefully matched on observables, which could have been in the guideline but were not. Following inclusion in a guideline, subfields grow in size and scientific impact. Rather than the aperture of subsequent innovation narrowing, subfields shift towards exploration as they become more translational, more intellectually distant, more disruptive, and build on more diverse and less established prior research, and this shift is larger for more technologically immature subfields. This demonstrates how knowledge certifying institutions may facilitate the use of existing knowledge as an input in the production of new knowledge, and how an institution designed to enhance the quality and decrease the cost of care can have an inadvertent but large impact on innovation. As certifying knowledge requires the investments of resource and time, this also speaks to how commitment can facilitate breakthrough innovation.

In the second essay, “Long-term Effects from Early Exposure to Research: Evidence from the NIH “Yellow Berets,” my coauthors and I ask if a relative short but intense exposure to frontier research can alter the career trajectories of potential innovators. To answer this question, we compile detailed career histories and innovation outcomes for applicants to the Associate Training Programs at the National Institutes of Health (NIH). We compare attendees to those who passed a first admission screen but were ultimately not selected, and after careful selection on observables, find attendees entered into—and remained in—research positions at higher rates and had greater research productivity as measured by publications, citations, patents, and grants. As there was only a small difference in the overall length of training, this speaks to how these results reflect the *content* of training rather than merely the fact that training was received. We also develop a set of measures to detect a “translational” style of biomedical research which was characteristic of the NIH at the time, and find attendees were durably imprinted with this style of research. Taken together, this study provides direct evidence that exposure to research can shift individuals into research careers, and shows how institutional workforce training efforts can alter the supply of innovators. It also speaks to how the topics studied by researchers reflects the organizational environment in which they were trained.

In the third essay, “Does Grant Peer Review Penalize Scientific Risk Taking? Evidence from the NIH,” I study a specific institution: grant peer review. This is motivated by a simple concern: while risk is an inherent aspect of innovation, those projects with high degrees of risk may be more likely to lead to breakthrough innovation yet may face challenges in winning the support necessary to be carried out. A fundamental challenge to empirically measuring any potential risk-taking penalty is that there is no

consensus on what constitutes risk in science nor how it should be measured. I discuss several types of scientific risk, including technical risk, execution risk, competitive risk, and career risk, each of which may have distinct dynamics. To address the challenging of measuring risk taking, I employ four separate measures—extreme tail outcomes, disruption, pivoting, and standing out from the crowd—each of which captures a distinct aspect of scientific risk taking. Using NIH R01 grant renewals after careful selection on observables, I find that grants with low levels of risk taking are renewed at higher rates, and conversely, grants with high levels of risk taking are renewed at lower rates. This provides direct empirical evidence that peer review at NIH punishes risk taking, confirming long-held anecdotal reports. Additionally, as a larger risk-taking penalty is seen for highly novel work compared to work with only marginal novelty, this argues the current degree of risk penalization results in a welfare loss. While I do not observe the mechanisms directly responsible for this risk penalty, I argue that, rather than being considered as an afterthought or as simply words of admonition to project evaluators, how to support the desired level of risk taking should be built into the design of supporting processes and institutions.

I conclude by highlighting some open research question. How does a physician’s clinical work shape their research, and conversely, how does their research work impact their clinical care? While it has long been claimed there is a connection between the two, it remains an empirically understudied phenomenon. In addition to being of direct managerial and policy interest for healthcare delivery and sponsors of biomedical research, this can help address questions related to human capital specialization and fungibility, as well as how problem exposure can shape research project selection. Additionally, more work remains to delineate the mechanisms involved in the essays above. Building on the first essay, how much of the observed treatment effect is due purely to knowledge verification as opposed to the other potential mechanisms discussed? This may deepen our understanding of the gap between the diffusion of codified knowledge and its effective deployment in the production of new knowledge. The second essay raises questions about the optimal timing of training interventions—as too early and individuals may not have the human capital to fully take advantage of it and it may separate exposure from the point of ultimate career decision, and too late and careers may no longer be as malleable—as well as how much weakening in terms of the quality and length of the training exposure is allowed before the treatment effect is substantially reduced. Finally, it is unclear in the third essay the relative contribution of the many potential mechanisms underlying the observed risk taking penalty—and by extension which potential managerial and policy solutions are most likely to be effective in mitigating it. I expect these and other questions using biomedicine to explore issues of broad interest within the strategic management of innovation to be a fertile area for future research, and one to which I look forward to making contributions.

References

- Azoulay, Pierre, Joshua S. Graff Zivin, and Gustavo Manso. 2011. "Incentives and creativity: evidence from the academic life sciences." *RAND Journal of Economics*, 42(3):527-554.
- Budish, Eric, Benjamin N. Roin, and Heidi Williams. 2015. "Do Firms Underinvest in Long-Term Research? Evidence from Cancer Clinical Trials." *American Economic Review*, 105(7):2044-2085.
- Catalini, Christian. 2018. "Microgeography and the Direction of Inventive Activity." *Management Science*, 64(9):3971-4470.
- Furman, Jeffrey L. and Scott Stern. 2011. "Climbing atop the Shoulders of Giants: The Impact of Institutions on Cumulative Research." *American Economic Review*, 101(5):1933-1963.
- Furman, Jeffrey L. and Florenta Teodoridis. 2020. "Automation, Research Technology, and Researchers' Trajectories: Evidence from Computer Science and Electrical Engineering." *Organization Science*, 31(2): 330-354.
- Graff Zivin, Joshua, and Elizabeth Lyons. 2021. "The Effects of Prize Structures on Innovative Performance." *AEA Papers and Proceedings*, 111:577-581.
- Kim, Soomi. 2022. "Shortcuts to Innovation: The Use of Analogies in Knowledge Production." Working paper.
- Krieger, Joshua, Danielle Li, and Dimitris Papanikolaou. 2022. "Missing Novelty in Drug Development." *The Review of Financial Studies*, 35(2):636-679.
- Lee, Sunkee. 2019. "Learning-by-moving: Can Reconfiguring Spatial Proximity Between Organizational Members Promote Individual-level Exploration?" *Organization Science*, 30(3):467-488.
- Lee, You-Na, John P. Walsh, and Jian Wang. 2015. "Creativity in scientific teams: Unpacking novelty and impact." *Research Policy*, 44(3):684-697.
- Murray, Fiona, Philippe Aghion, Mathias Dewatripont, Julian Kolev, and Scott Stern. 2016. "Of Mice and Academics: Examining the Effect of Openness on Innovation." *American Economic Journal: Economic Policy*, 8(1):212-252.
- Nagaraj, Abhishek, Esther Shears, and Mathijs de Vaan. 2020. "Improving data access democratizes and diversifies science." *Proceedings of the National Academy of Sciences*, 117(38):23490-23498.
- Tian, Xuan and Tracy Yue Wang. 2014. "Tolerance for Failure and Corporate Innovation." *Review of Financial Studies*, 27(1):211-255.
- Wu, Lingfei, Dashun Wang, and James A. Evans. 2019. "Large teams develop and small teams disrupt science and technology." *Nature*, 566(7744):378-382.

Chapter 2

Anchoring Exploration: Knowledge Certification and the Dynamics of Innovation

Abstract

How does knowledge certification dynamically shape the rate and direction of innovation? To empirically explore the role of knowledge certification—which may act to decrease the static and dynamic uncertainty of building upon extant knowledge—I study how 135 clinical practice guidelines published 1990-2011 by cardiology professional medical societies shape the use of the 14,390 underlying articles and associated research subfields they incorporate. Employing a difference-in-differences approach, I find subfields included in the guideline have a large and sustained increase in the flow of articles into the field, which are more likely to be themselves highly cited, compared to controls carefully matched on observables. Rather than the aperture of subsequent innovation narrowing, subfields shift towards exploration as they become more translational, more intellectually distant, more disruptive, and build on more diverse and less established prior research, with only a small increase in incremental clinical research. In particular, this shift is larger for those subfields that are more technologically immature. As certifying knowledge requires the investment of resources and time, this shows how commitment can facilitate breakthrough innovation.

1. Introduction

The first β -blockers were developed in the early 1960's, work contributing to Sir James Black earning a Nobel Prize in Physiology or Medicine in 1988. Initially, physicians were reluctant to test β -blockers in heart failure out of concern their negative effects on heart contractility might be harmful (Funck-Brentano 2006). By the mid-1970's, observational reports started to trickle in showing a benefit to the drugs in heart failure, however, many physicians didn't believe these results and researchers were similarly reluctant to study them. This uncertainty led to a substantial delay in innovation: it wasn't until 1993 that the first randomized, placebo-controlled trial showed a survival benefit. It was another six years until the Heart Failure Society of America reviewed and synthesized the literature, certifying their recommendations in a guideline for the use of β -blockers in heart failure (Heart Failure Society of America 2000). There was a resulting flurry in follow-on innovation, including efforts to develop novel mechanisms targeting the β -blocker pathway.

The history of β -blockers in heart failure illustrates the central challenge of this paper: How do knowledge workers decide which knowledge to build on and in which direction to steer the resulting innovation? A key element of the above example is that little was stopping researchers from *choosing* to study β -blockers in heart failure. The key innovation, that β -blockers should be started with a low dose and titrated upwards, was simple and available for study, and yet took decades until it was rigorously tested. Similarly, from a technical standpoint, much of the innovation researchers pursued after the guideline could have been achieved beforehand, as it built entirely on knowledge that already existed. Rather, an important reason researchers chose not to build on it is uncertainty: they didn't view these early finding of β -blockers in heart failures as a reliable basis for follow-on innovation.

More generally, a key input in the production of new knowledge is old knowledge (Scotchmer 1991). Yet, not all extant knowledge is available as an input for the production of new knowledge (Mokyr 2002) or ensures its use in the development of new inventions (Nelson 1959). Knowledge must be codified so it is readily transmissible, have diffused so a user is aware of it, and a would-be-innovator may need to pay a transaction cost to access and internalize the knowledge (Foray 2004). Innovators must also have the capabilities and complementary assets that allow successful application of the knowledge (Cohen and Levinthal 1990).

This approach, however, overlooks another central reason why extant knowledge may not be used in the production of new knowledge: uncertainty in the knowledge itself. This uncertainty occurs along two dimensions. First, there is static uncertainty in the validity of any single piece of knowledge. The history of science and technology is filled with prominent findings initially not used as a basis for cumulative innovation because they were widely not believed. Examples include Ignaz Semmelweis and handwashing preventing postpartum infections (1847), Barry Marshal and Robin Warren and *H. pylori* as the cause of peptic ulcers (1982), and William Coley and cancer immunotherapy (1891). This problem may be especially

acute for the knowledge generated by women, minorities, and low status researchers, whose work may systematically be viewed with more uncertainty independent of its quality (Bikard and Fernandez-Mateo 2022, Koffi and Marx 2022).

There is also dynamic uncertainty. It is often unclear what development trajectory a technology will take as well as what complementary innovations may be developed (Rosenberg 1995, 1976). In addition, it is frequently difficult to predict the ultimate commercial and status returns for a successful innovation. Given the long time frames involved in some innovative efforts, this makes it challenging for researchers today to select the project which in the future will lead to the largest professional returns when completed.

By shaping the perceived and real risk and returns to pursuing particular research and development projects, shifts in the degree of static and dynamic uncertainty in knowledge may influence which projects researchers choose, how they pursue them, and which development trajectory they take (Foster et al. 2015, Franzoni and Stephan 2021). In so doing, efforts to certify knowledge, which involves actors with authority evaluating and synthesizing across extant knowledge and may act to reduce the uncertainty of follow-on innovation, may increase the effective supply of useful knowledge available for cumulative innovation.

While knowledge certification might increase the overall supply of useful knowledge, its potential effect on the direction of innovation is less clear, in particular, whether there is a shift towards exploration or exploitation following certification (March 1991). A concern among centralized organizational and institutional efforts to support innovation is that they may constrain the creative search process and shift follow-on research towards incremental advances and away from radical innovation. Scholars have raised this concern for related activities such as process management (Benner and Tushman 2003), standards (Farrell and Saloner 1985), formal coordination mechanisms (Jansen et al. 2006), knowledge management systems (Lee and Van den Steen 2010), and codification (Garcia-Muina et al. 2009). The impact of a shift towards incremental research early in the course of technology’s development—precisely the stage with the most static and dynamic uncertainty—may be especially concerning, as failure to explore promising potential development paths may lead to lower ultimate technology performance (Gans et al. 2020). On the other hand, if certifying knowledge reduces uncertainty more for exploratory rather than exploitative work, follow-on innovation may shift towards exploration. Additionally, some scholars have also argued problem structure and constraints may facilitate creativity in some settings (Ehls et al 2021, Rosso 2014, Stokes 2005).

Empirically, estimating the impact of knowledge certification on both the degree and type of innovation produced from existing knowledge is difficult. First, in some settings knowledge is hard to observe and characterize transparently. Knowledge may be tacit, and may not easily be directly comparable to other knowledge. Second, any shock must only apply to some rather than all knowledge within a domain to be able to develop a counterfactual. Settings with winner-take-all markets, such as technology standards, or other sharp constraints on innovative activity can make it difficult to credibly measure unaffected knowledge. Finally, changes to underlying knowledge characteristics are often endogenous, limiting causal

interpretation. In line with these challenges, prior empirical literature on how knowledge characteristics impact innovation is limited. Studies have looked at closely related processes, such as codification, but these have largely been case studies or descriptive (Balconi 2002, Garcia-Muina et al. 2009, Lissoni 2001). I will argue clinical practice guidelines provide an empirical window useful to overcome these challenges.

Clinical practice guidelines (CPGs) produced by medical professional societies evaluate and synthesize the medical literature, providing practical recommendations to guide clinical management decisions around disease diagnosis and treatment. CPGs are typically developed through an intensive, formal process where a panel of respected experts, selected by their peers, is convened to review the literature on behalf of the professional society with deliberate procedures in place to arrive at consensus. They frequently will explicitly rate both the importance to patient care and the quality of the evidence supporting their recommendations. In this way, CPGs certify the knowledge they contain and act to reduce the static and dynamic uncertainty of building on it. Physicians still have bounded autonomy in the care of patients and research to depart from the CPGs (Mello 2001). By typically being published in academic journals, CPGs allow for a rich set of bibliometric outcomes by which to characterize follow-on research.

The Agency for Healthcare Research and Quality (AHRQ) was established in 1989 under the U.S. Department of Health and Human Services to improve the effectiveness and appropriateness of health care delivery. Its initial mandate included facilitating the production of CPGs, which rapidly rose in number in the following years. Crucially for my purposes, the development of CPGs was driven by operational concerns to decrease cost and improve quality of direct clinical care delivery, but had no consideration for any potential impact on innovation (Weisz et al. 2007, Woolf 1990). While which topics received a guideline is endogenous, the exact timing may be considered exogenous.

I study the extant knowledge embodied in 14,930 original research articles cited by the 135 CPGs published in PubMed by professional cardiology associations in the United States, United Kingdom, Canada, and Australia between 1990 and 2011. To identify the effect on follow-on innovation of inclusion of each paper in a CPG, I must estimate the counterfactual outcome of what would have happened had each paper not been cited by a CPG. Rather than focusing on citations to trace knowledge flows over time, I capitalize on the PubMed Related Citations Algorithm to delineate subfields of research articles based on shared intellectual content constructed using the textual similarity between articles in their titles, abstracts and keywords assigned by the National Library of Medicine. This allows me to follow how a *subfield* evolves in response to knowledge certification (Azoulay et al. 2019). I employ a difference-in-differences estimator to study the within-subfield change in follow-on research characteristics around the time of inclusion in the CPG compared to a set of cardiovascular (CV) papers and their associated subfields carefully matched on observables which could have been included in the CPG but were not. Given that potential control articles closely related to treated articles may be negatively affected by their exclusion from a CPG, as well as that distantly related potential controls may not have a parallel trajectory, I explore these issues with a number of alternative control groups.

I find subfields included in a CPG have a large and sustained increase in research and impact. There is a 9.5% increase in the flow of articles into a subfield following the CPG, and papers within the subfield become more impactful as measured by citations. I also find evidence the direction of research in a subfield durably shifts. While there is an increase in incremental research, there is a much larger increase in exploratory research. Research within the subfield becomes more translational, draws from more diverse and less established prior research, is more intellectual distant from the knowledge it builds on, and is more disruptive following the CPG. There is also a greater commercial orientation within the subfield, with a larger increase in articles cited by a patent. This shift towards exploration and increase in follow-on research is more pronounced for more technologically immature subfields than more established subfields. While multiple mechanisms are likely to play a role, I provide evidence this effect is not primarily driven by an increase in attention, codification, problem prominence by changing provider practice, or social influence.

This study’s primary contribution is to show direct evidence that, rather than narrow the aperture of follow-on research, knowledge certification produces an increase in more exploratory research. In line with this, it suggests a practical tool for managers of innovation and policy makers in stimulating radical innovation is commitment (Gans et al. 2020). Rather than rapid, low-cost experimentation, radical innovation may be better supported by making the investments necessary to de-risk key assumptions and the expected development trajectory for a particular technology. Similar to Finkelstein’s study of policies to increase usage of preexisting vaccines (2004), the dynamic effects of CPGs on innovation are large and unintended. This shows how an institution aimed at professional practice can have a substantial impact on innovation, including providing systematic evidence of how applied science can inform basic science (Gelijns and Rosenberg 1994, Stokes 1997). Knowledge certification may have large positive externalities, making it undersupplied by market mechanisms and potentially benefiting from specific institutions to support their development. It also may provide a partial solution to the challenges of learning the ever-increasing stock of knowledge to reach the scientific frontier, as innovators may rely on the judgment of certifiers rather than needing to synthesize and evaluate this prior knowledge themselves (Jones 2009). Finally, it emphasizes the value of considering the state of technological maturity in the study of innovation, as mature and immature technologies may respond very differently to changes in their organizational and institutional environments.

The rest of the paper proceeds as follows. In Section II, I develop the underlying theoretical background to this study before discussing in the institutional context in Section III. Section IV presents the data, empirical strategy, and descriptive statistics. I report results in Section V. Section VI concludes with the implication of this study.

2. Knowledge certification and innovation

Producing knowledge available for use in cumulative innovation can be conceived of as a system (Foray 2004, Mokyr 2002). Knowledge, initially tacit, needs to be codified so it can be readily transmitted

(Polanyi 1966, von Hippel 1994). This codified knowledge must then diffuse so other researchers are aware of its existence (Sorenson and Fleming 2004). Finally, would-be innovators may need to pay transactions costs to both directly access and internalize this knowledge (Murray et al. 2016). In each of these steps—generating tacit knowledge, codification, diffusion, and access—there may be a significant gap between the private and social returns of associated investments. The incentives and frictions between these stages, and hence the amount of useful knowledge available for cumulative innovation, is endogenously shaped by supporting institutions.

Yet, merely being aware and having access to codified knowledge, as well as having the human capital and complementary assets necessary to build on it, does not guarantee knowledge workers will see it as a useful basis for further innovation. One reason for this gap between access and use is uncertainty. This occurs both from static uncertainty in the reliability of the knowledge itself, and dynamic uncertainty in the future trajectory and returns to research. While scientific fraud may be a high profile source of static uncertainty, more commonly, uncertainty stems from the quality of the science itself and to what extent it reinforces or conflicts with scientists’ prior beliefs (Azoulay et al. 2015). Importantly, determining whether knowledge is valid is not a costless activity. It requires investments of time, expertise, and in some cases direct resources. The difficulty in evaluating knowledge validity is exacerbated by concerns a significant amount of published science is neither reliable nor reproducible (Begley and Ioannidis 2015, Ioannidis 2005); institutional norms which do not reward replication (Dasgupta and David 1994); information asymmetry between the original innovator and those evaluating the work, and the sheer volume of scientific output where only a small fraction of article, even in one’s own field, can be read, let alone rigorously evaluated.¹ In addition to static uncertainty, there is also dynamic uncertainty: it is often unclear which development trajectory a technology will take; what complementary innovations may be developed; and what the potential economic, status, and career rewards will be in the future when deciding today which innovation projects to pursue (Rosenberg 1995, 1976). In many cases the degree of static and dynamic uncertainty will be larger for early-stage technologies than their more mature technologies,

How might knowledge certification reduce static and dynamic uncertainty? Knowledge certification may shape innovation more generally through a number of channels. This includes codifying otherwise disparate or tacit knowledge sources and facilitating knowledge diffusion, which act to reduce the search and transaction cost of accessing the knowledge, or in the case of certification of professional practice, altering problem prominence by changing professional behavior (Rosenberg 1969). In particular, however, three potential mechanisms of certification may shape the amount of static and dynamic uncertainty: knowledge evaluation, knowledge synthesis, and coordination.

¹ The scope of this problem is easy to illustrate. In 2011, the last year of the present study, 45,439 original research articles about CV diseases were published in PubMed, a 254% increase since 1990. A cardiologist devoting one afternoon a week to keeping up on the literature, assuming one article read per hour, would be able to read under 1% of published research in her field. This concern is present even in more narrow fields: in 2011, on average, 15 original research articles are published on heart failure each day.

Certification may reduce the uncertainty in building off knowledge through direct evaluation of the knowledge. Consider two illustrative examples. First, Judah Folkman’s initial findings on the role of angiogenesis in cancer were published in 1971, and yet for reasons of technical uncertainty in how the experiments were interpreted, for many years few believed his findings and almost no other scientists build on this work. Second, the Wright brothers’ first flight occurred in December 1903 in Kitty Hawk. Yet few believed the Wright brothers had flown until a large-scale public demonstration in France in 1908, and in the intervening time, other aviators were not building on and improving the Wright brother’s design. Here, it was quite straightforward to prove if the Wright brothers had actually flown. Rather, other innovators did not expend the time and resources necessary to evaluate if their claim of a successful flight was valid. In both cases, rigorous evaluation would have been able to reduce the associated uncertainty and speed follow-on innovation. While this most directly speaks to static uncertainty, evaluation may also change the expectations for future follow-on innovation.

A second closely related idea is that additional information—with an associated reduction in uncertainty—is available from synthesis across knowledge. Knowledge is often conceptualize as discrete knowledge objects, which is reflected in how it is frequently measured in counts of individual patents, publications, and products (Foray 2004). In fact, however, not only is knowledge cumulative, it is also highly interdependent. The validity of information contained in an individual piece of knowledge is not independent, in a statistical sense, from the validity of closely related knowledge in the same field. Consider how many elements true for influenza were true for COVID-19 as well because they came from the same fundamental biologic data generating process. This connection is often not reflect in the direct citations so frequently used in the study of innovation. What’s more, additional information can be contained in the synthesis from multiple extant pieces of individual knowledge. This can very directly be seen in how knowledge uncertainty is lower after synthesis within a meta-analysis than it is in the individual source articles that inform the meta-analysis. Integration across knowledge can also have an emergent phenomenon, where the impact and understanding of the whole is greater than the sum of its parts as higher-level concepts and processes become apparent. Examples include observations around the role of the immune system in modulating cancer and critical periods in neuronal plasticity.

Finally, knowledge certification may act as a coordinating mechanism across researchers, reducing dynamic uncertainty by altering the expectation of what peer researchers will be studying and what the future demand will be for any resulting innovations. This may be shaped both by the status and authority of the certifying institution or agent, but also by the quality of the certification process. Coordination may occur both through push-mechanisms, most closely related to the status and authority of the certifying institution or agent, with direct efforts to alter or constrain researcher behavior, and by indirect pull-mechanisms related to the expected returns to innovation. In the case where there are sharp constraints on innovator activity, certification begins to more closely resemble standards and may have distinct concerns (David and Greenstein 1990, Farrell and Saloner 1985, Rysman and Simcoe 2008).

Three factors about knowledge certification are essential to emphasize. First, the degree of knowledge certification is a spectrum rather than having a sharp boundary between certified and not certified knowledge. For instance, case reports, case-control studies, and randomized controlled trials are all published in peer-reviewed medical journals, and to some extent are certified knowledge, but all have different levels of remaining uncertainty about their findings (Polidoro 2013). Second, this implies the impact of knowledge certification depends not just on the status of an institutional body or actor facilitating it, but also on the quality and rigor of the certification process. The potential reduction in uncertainty for a high-status body simply anointing knowledge as certified without accompany rigorous evaluation and synthesis of this knowledge may be very different. Similarly, distinct dynamics may be present for the knowledge included in review articles, which may have evaluation and synthesis without high-status or institutional endorsement (McMahan and McFarland 2021, Salandra et al. 2021). Finally, to the extent knowledge certification can require substantial time, effort, and direct resources to undertake, it may be viewed as a form of commitment.

How might certification shape the amount and direction of innovation by reducing the static and dynamic uncertainty of building on certified knowledge? Uncertainty is an important consideration by knowledge workers when selecting research projects (Foster et al. 2015, Franzoni and Stephan 2021). Researchers may be motivated by job security, reputation, and funding concerns, where the value (risk) of a successful (failed) project may advance (derail) career progression (Stephan 2012). These concerns will be distinct across different stages and loci of a career, with on average, different risk-reward preferences for junior versus senior researchers. In short, knowledge certification changes the expected value, both in terms of pecuniary and non-pecuniary rewards, of research building on this knowledge. This can change not only what researchers chose to study, but also how much resources they devote to it and the direction they steer the innovation.

These same issues are also at play at firm-level responses to new knowledge. Firms may want to build the corresponding absorptive capacity (Cohen and Levinthal 1990), switch to a new technology (Christensen 1997), or alter informational filters within the firm (Henderson and Clark 1990). However, given uncertainty about its validity and the expected course of follow-on innovations, as well as the costs of these responses, it is difficult for managers to know to which new knowledge their organization should respond and how they should respond (Gans 2016).

While certification might make more knowledge effectively available for cumulative innovation, there is a concern it might shift follow-on work towards incremental innovation. In particular, it might constrain innovation to local, bounded search around the certified knowledge or lead to greater homogeneity in the direction of technological development (Mihm et al. 2010, Stuart and Podolny 1996). If this were true, knowledge certification early in the course of a technology’s trajectory might actually be ultimately detrimental as potentially promising technology development trajectories may remain unexplored, hurting ultimate technological performance (Gans et al. 2020).

Several potential mechanisms, however, might act to countervail this. First, knowledge certification may not alter the expected value of all potential research projects equally. While ultimately an empirical question, it seems likely the reduction in uncertainty would be greater for highly uncertain rather than highly certain areas of research. As an extreme example, certification of the value of physician handwashing is likely to produce only marginal change in the expected value of follow-on handwashing research, while doing so for mRNA vaccine technology prior to the COVID-19 pandemic may have a larger impact on what researchers view as promising avenues for study. Notably, the effect on research project selection acts through changing the *perception* of risk, which may systematically depart from the true underlying risk of a project (Tversky and Kahneman 1973). Second, certified knowledge ties into competitive dynamics within scientific research (Hill and Stein 2021). By raising the bar of what is considered a contribution to the field, the knowledge certification may push researchers to choose more risky projects to maintain the same scientific status relative to their peers. Risk compensation provides another link between uncertainty and the selection of research projects.² As the overall uncertainty in particular projects decreases, researchers may compensate by feeling comfortable selecting (what were previously) more risky projects. Together, these conflicting processes make it difficult to predict the relative impact of knowledge certification on incremental versus radical innovation.

3. Clinical practice guidelines

The Institute of Medicine defines a CPG as, “statements that include recommendations intended to optimize patient care. They are informed by a systematic review of evidence and an assessment of the benefits and harms of alternative care options” (2011). While the focus is on providing practical guidance for clinical management decisions, typically containing a set of enumerated recommendations to guide healthcare providers, they combine this with a discussion of the published literature underlying the recommendations including its strengths and limitations. In this respect, they may be seen as containing both fundamental and applied knowledge (Stokes 1997). A unique aspect of CPGs is many of them explicitly rate the degree of benefit and level of evidence supporting their recommendations.

While there is variation across professional societies in the processes used to construct CPGs, these are typically very intensive processes. Over the course of 1-2 years, a committee of leaders in the field systematically evaluates and synthesizes the evidence, often with formal consensus mechanisms in place. In this way, CPGs can be viewed as shifting the knowledge they include in the direction of being certified. Appendix B provides additional details and examples of the contents of CPGs and processes used to construct them.

² Risk compensation, also known as the “Peltzman effect,” grew out of the study of automobile safety regulation. Samuel Peltzman (1975) argued—controversially—that federally mandated safety regulations, such as seat belts for all occupants, did not reduce highway fatalities as the safety improvements were completely offset by an increase in risky driving behavior.

History and evolution

CPGs are not a new phenomenon.³ However, they tended to be infrequent and ad hoc until the late 1980's (Weisz et al. 2007, Woolf 1990). In the 1980's three observations were central to the advent of CPGs (Brook 1989). First, the rate of increase in healthcare costs was much higher than the growth of the broader economy, pushing the government and payers to look for solutions to decrease the growth of healthcare costs (IOM 1989). Second, a stream of research noted widely different medical practices across the country, where equivalent patients would get very different treatments based upon the specific provider seen, along with some delivered health care services determined to be of little or no value (Chassin et al. 1987, Hughes et al. 1984, Rosenthal 2012, Winslow et al. 1988). Third, many areas of medicine had relative little evidence supporting their practice, making it difficult to draw firm conclusions on proper care (Eddy and Billings 1988). Motivated by perceptions of inappropriate care, this stimulated a movement in medicine tightly linked to the evidence-based medicine efforts of the 1990's and rise of "delivery science" that sought to improve the implementation of medical practices guided by underlying data. Together, these forces led to a push for guidelines to improve quality and decrease cost (IOM 1992).

The U.S. government convened commissions to study this, ultimately resulting in the 1989 establishment under Public Law 101-239 of the Agency for Healthcare Research and Quality (AHRQ).⁴ The AHRQ was tasked with activities that will "enhance the quality, appropriateness and effectiveness of health care services" with a mandate to develop, evaluate, and disseminate CPGs, providing a "shock" to guideline production (IOM 1990, Woolf 1990). Rather than produce guidelines themselves, the AHRQ arranged joint public-private efforts to develop CPGs through contracts with private groups or convening nongovernmental expert panels.

Prior research has carefully documented the impetus and role of professional medical societies in CPGs, upon which I lean heavily here (IOM 1992). Initially in an effort to preserve their autonomy and ward off potential regulation and administrative pressure, including more stringent controls from Medicare payments, medical professional societies sought to self-regulate by developing their own CPGs. This effort would allow decisions to be made by professional peers developing the CPGs rather than government bureaucrats, and facilitate a primary focus on quality improvement rather than cost containment (IOM 1992, Woolf 1990). This is supported by a survey from the U.S. General Accounting Office (1991) of professional medical societies, who stated the two primary motivations for engaging in the production of CPGs were to increase the quality of care and to defend the specialty against outside forces including

³ For instance, the American Academy of Pediatrics-sponsored *Red Book: Report on the Committee on Infectious Diseases*, which includes guidelines on the treatment of infectious disease, dates to a 1938 pamphlet "Immunization procedures." Similarly, the American College of Obstetricians and Gynecologists published its first practice guidelines in 1959.

⁴ At the time of founding, the agency was called the Agency for Health Care Policy and Research, which was renamed to AHRQ under the Healthcare Research and Quality Act of 1999.

regulators, payers, and malpractice litigants. Reflecting these trends, at the annual meeting of the Council of Medical Specialty Societies in 1987 the society’s president summarized the meeting that it is the “duty of specialty societies and physicians to lead in shaping quality guidelines” (Council of Medical Specialty Societies 1987). Professional societies responded by creating committees to oversee the development of guidelines and adopting formal procedures for doing so (Woolf 1990). As noted by the Institute of Medicine (1992), “Overall, professional societies are generally seen as having taken the lead in organized efforts to develop practice guidelines, thereby serving as an important, perhaps key, source of authoritative communication to practitioners about what constitutes appropriate care.”⁵

The increasing role of CPGs in medicine is born out in the data: in 1990, 20 CPGs were published by U.S. professional medical societies, while in 2018 the number had grown to 115. From 1990-2018, a total of 3,269 CPGs were published by English-language medical professional societies in academic journals. The increasing prominence of CPGs is perhaps best emphasized by the institutional structures that have developed to support them. There are now numerous “guidelines on how to develop guidelines” in medicine (Mechanick et al. 2017), with many professional organizations sponsoring standards on the production of standards (IOM 2011). Similarly, there are now systematic reviews on how evidence-based the evidence-based guidelines are (Lee and Vielemeyer 2011, McAlister et al. 2007, Tricoci et al. 2009), impact on patient care (Grimshaw et al. 1995, Lugtenberg 2009), quality of guideline processes (Jue et al. 2019), adherence (Ebben et al. 2013), and tools to increase adherence (Flodgren et al. 2016). A large literature has documented the impact of CPGs on clinical care, including that increased adherence is associated with better clinical outcomes and reduced costs (Komajda et al. 2005, Wilke et al. 2011, Wockel et al. 2010).

CPGs and the study of innovation

CPGs provide an empirically attractive setting to understand the impact of knowledge certification on innovation. Given the well-documented motivations around cost and quality of care, the implementation of CPGs was not driven by innovation concerns. The most common motivations for selecting which topics received CPGs include disease prevalence, degree of mortality and morbidity, cost of managing the disease, and the potential for a CPG to improve these three categories (IOM 1995). While which topics receive a CPG is endogenous, in practice, the exact timing may be considered exogenous. First, while there is variation in the process used, the creation of a CPG is often a multi-year affair, temporally separating any preceding factors stimulating CPG production from CPG release and measurement of its impact (Woolf 1992). Second, professional societies, especially in the early years of CPG production, had a tendency to work on one or two CPGs at a time. This led to an important reason for the timing of a particular CPG to be the limited resources dedicated to CPG writing more generally, requiring their sequencing. The level of interest, as well as capacity for CPG writing, varied idiosyncratically across professional societies.

⁵ Appendix B discusses CPGs produced by organizations other than the professional medical societies.

Additionally, CPGs are frequently published in academic medical journals, allowing for a rich set of bibliometric characteristics and outcomes for analysis. Professional norms lead most medical innovations to be published in peer-reviewed journals, with patents playing an important role in protecting biopharmaceutical and medical device innovations (Dasgupta and David 1994). Due to this, I can follow a robust “paper trail” of innovation, which may not be present in fields where trade secrets play a larger role. While practitioners have raised concerns CPGs may alternatively stimulate or inhibit medical innovation (Cerfolio 2018, Law 2017, McCann 2020, Pandyan et al. 2013), to my knowledge no prior research has systematically examined the impact of CPGs on innovation.

CPGs are also attractive from a theoretical perspective. In contrast to the study of winner-take-all (or most) markets, healthcare providers still have bounded autonomy in both the care of patients and research to depart from CPGs, which typically leave significant room for physician discretion (Blind 2004, David and Greenstein 1990, Mello 2001). Unlike research on other institutions impacting innovation, such as the patent system (Scotchmer 1991, Williams 2013) or repositories (Furman and Stern 2011), CPGs may have lowered search costs but did not directly affect access to underlying inputs for innovation. Additionally, in contrast institutions deliberately built to stimulate innovation, any impact of CPGs on innovation was unintentional. While it is difficult to preclude any role for strategic efforts by firms to direct the CPG creation process to enhance their ability to capture value, the strong professional norms and frequent explicit requirements to declare any conflicts of interest suggests this is not a first-order concern.

4. Empirical design, data and descriptive statistics

Data and measurement

Clinical practice guidelines

I begin by identifying all CPGs published by CV professional societies starting in 1990 following the creation of the AHRQ. The empiric focus on CV CPGs has multiple benefits. First, CV is a large field of medicine with a relatively high level of evidence supporting its practice. Additionally, CPGs have a prominent role within the specialty. Finally, empirically, this focus facilitates a sufficient granularity of measurement and manual collection of data where required. These CPGs were manually identified from CV professional society webpages. To ensure the capture of old or retired CPGs, I also carefully reviewed CPGs for any citations to prior CPGs, as well as searched through PubMed, an online resource from the National Library of Medicine which indexes, as of 2019, over 30,000 journals and nearly 30 million articles in the biomedical sciences. Given the possibility the effect of a CPG may only be observed in settings bounded by its institutional regime, I focused only on CPGs published in English by professional societies in the United States, Canada, United Kingdom, Australia, and New Zealand. This excludes CPGs by pan-European professional medical societies where the institutional regime supporting a CPG in a particular locale is less clear, as well as English language CPGs by countries whose primary language is not English.

As bibliometrics and the ability to trace subfields and knowledge flows is central to this study, the analysis was limited to CPGs published in peer-reviewed journals indexed in both PubMed and Web of Science, which contains citation data. Since Web of Science data was only available through 2014, I only include CPGs through 2011 to allow for three years of subsequent citation data. CPGs with multiple chapters published as separate articles in PubMed are considered as a single CPG. This produces a final sample of 135 CPGs.

Table 1a presents CPG descriptive statistics. Reflecting that CPGs are frequently significant, intensive efforts to construct, the average CPGs has 20 authors (range 0-63), is 45 pages long (range 3-260) and contains 243 references (range 1-1,194). CPGs may simultaneously be published in multiple journals—typically journals closely associated with each sponsoring professional society—resulting in, on average, each CPG being published in 1.8 journals (range 1-6). CPGs in the sample are dominated by those from the United States and Canada, with 90% of CPGs having at least one sponsoring professional organization in United States or Canada, and 6.7% of CPGs having sponsoring professional societies from multiple nationalities.⁶ Both the number and extent of guidelines has been increasing over time (Appendix Figure A1). While in 1990-1992 there were 7 CV CPGs with an average of 6.9 authors and 67 references, in 2009-2011 there were 46 CV CPGs with an average of 20.9 authors and 220 references.

CPGs are not review articles. They typically go through a more extensive literature review, are developed using a formal drafting process with protocols in place, have a much larger and typically very prominent authorship, are associated with more authority by virtue of the professional society approval, and are much more visible within the profession. Additionally, CPGs frequently offer formal, explicit recommendations for practice and the degree of supporting evidence, rather than simply summarizing the literature. This difference is reflected in the descriptives (Appendix Table A1). Compared to the CV review articles published during the same time period, the mean CPG has 7-fold the number of authors, over 10-fold the citations, and nearly 5-fold the length and number of original research article references.

Professional societies may periodically update their CPGs (Table 1b). I term these guidelines published by the same society (or group of societies) on the same topic to be a guideline series. The 135 CPGs under study comprise 73 separate guideline series, with 53% of the guideline series ever receiving an update with an average length of time of 5.7 years between updates. Conversely, 75% of CPGs are part of a guideline series receiving an update during the time period studied.

Since I am interested in how knowledge certification shapes the use of extant knowledge and their associated subfields, I empirically focus on original research references within CPGs. For clarity, I term these references root articles. The focus on original research excludes any review articles or commentary.

⁶ There is some variation across nationalities and professional societies in the extent to which CPGs are published in academic journals. While most U.S.-based CPGs are published in journals, many British guidelines, such as those from the National Institute for Health and Care Excellence (NICE) are shared publically but not published in academic journals.

The 135 CPGs cite a total of 21,138 articles. I drop 82 articles not in English, 3,858 that are not original research articles, 18 for which no MeSH terms were available, and 2,250 without any CV disease MeSH, for a final sample of 14,930 root articles. Of these, 5,028 (33.6%) are cited by multiple CPGs; I will exploit this variation as a robustness check.

Identifying research subfields

To delineate research subfields, I capitalize on the PubMed Related Article (PMRA) algorithm. The algorithm, developed by Lin and Wilbur (2007), has previously been employed in the study of science to identify research subfields (Azoulay et al. 2015, Azoulay et al. 2019) and scientific similarity (Myers 2020).⁷ Briefly, the algorithm takes as inputs an article’s title, abstract and keywords assigned by the National Library of Medicine, and after weighting terms by their overall frequency within the database such that rare words hold a greater weight, identifies the set of articles that have the greatest textual similarity. For each article, the algorithm returns the 100 articles with the greatest textual similarity to the focal article, subject to a constraint of a minimum similarity threshold score shared across all articles. In addition to the 100 most similar articles, the algorithm expands the articles it returns by reciprocity: All articles considering a focal article to be among their most related articles are also considered by the focal article to be related. In all cases, I only include original research subfield articles. This excludes other types of publications such as review articles, letters, and erratum. In practice, subfield size can vary widely; the size of the subfields in my ultimate treated and control subfields ranges from 14 to 4,954 with a mean of 372. Appendix Figure C1 provides an example of a root article and its associated subfield.

A crucial aspect this approach is it does not rely on citations, precluding concerns about strategic, social, or signaling motivations for citation. Additionally, reliance on article text allows subfields to rapidly reflect changes in scientific direction. While authors may shape their word choice in response to a CPG, this approach to subfield delineation is based on research *content* (Thompson and Hanley 2017). As any idiosyncratic features of the PMRA are shared across both treated and control articles, the difference-in-differences empirical strategy resolves concern these are driving the results.

Research topic, methodology, and the direction of innovation

To characterize the content and methodology of articles, I capitalize on Medical Subject Headings (MeSH). MeSH is a hierarchical controlled vocabulary created and maintained by the National Library of Medicine used for indexing and searching the biomedical literature. In the 2019 edition of MeSH, there are over 29,000 terms, with an average of about 12 MeSH keywords associated with each PubMed publication. Crucially, MeSH are assigned by professional indexers. This excludes any concern authors selecting MeSH

⁷ Detailed description of the algorithm may be found at <https://ii.nlm.nih.gov/MTI/Details/related.shtml> as well as in Azoulay et al. (2019). Data was harvested using an open-source software tool available at <http://stellman-greene.com/FindRelated/>.

might do so strategically (Bachrach and Charen 1978). While it is difficult to assess accuracy as there is no single “correct” way to index articles, MeSH have generally been found to be consistent across indexers (Coletti and Bleich 2001, Funk and Reid 1983). MeSH have been previously used in economic studies to characterize biomedical research, and form much of the raw input I use to characterize the direction of innovation (Azoulay et al. 2021, Azoulay et al. 2010). A significant advantage of MeSH is the granularity and transparency they provide into what is being measured compared to alternative approaches such as machine learning textual analysis or recombination of patent classes.

I use MeSH terms, as well as titles, keywords and publication types, to develop four measures of research methodology: clinical trials, case reports, epidemiologic studies, and bench science. Clinical trials includes terms related to phase 1, 2, 3, or 4 clinical trials (e.g. *phase I clinical trial*). Similarly, case reports are tagged with the *case reports* publication type and epidemiologic studies contain MeSH such as *case-control studies* and *cohort studies*. Bench science is characterized by papers employing a molecular biology technique (e.g. *nucleic acid amplification techniques*), using a molecular biology model organism (e.g. *Caenorhabditis elegans*), or having terms related to cellular structures and macromolecules (e.g. *DNA topoisomerase IV*) or biochemical and cellular processes (e.g. *oxidative phosphorylation*) (Azoulay et al. 2021). Together, 53.6% of CPG root articles are characterized by one of these measures, with the remaining root articles largely comprised of observational clinical studies. I also measure “translational” research. While translational research does not have an agreed-upon definition (Butler 2008, Woolf 2008), it focuses on bridging the “bench to bedside” and “bedside to bench” transition. Following Azoulay et al. (2021), I consider translational research to use the tools of bench science to study disease, being tagged by at least one bench science keyword and one disease keyword. Full details on the construction each of these measures are available in Appendix C.

Author response

To identify author response, I rely on Author-ity for author disambiguation of PubMed articles (Torvik and Smalheiser 2021, Torvik and Smalheiser 2009, Torvik et al. 2005). Author-ity is driven by the insight that authors of the same name sharing coauthors, MeSH, affiliations, title keywords, and other such characteristics are likely, in fact, to be the same person. Author-ity has been shown to be highly accurate (Lerchenmueller and Sorenson 2016). I leverage the disambiguated authors to determine the number of new authors within a subfield, as well as to calculate measures of author experience and expertise based on their prior publications. A longstanding norm in biomedical research has the most senior author or principal investigator typically listed last, while the junior author primarily responsible for conducting the research

is listed first, with middle authors listed in order of their contribution (Dance 2012, Sauermann and Haeussler 2017). In line with this, my analysis focuses primarily on last authors.⁸

Empirical strategy

I am interested in how knowledge certification by CPGs shapes the amount and direction of innovation with subfields. To identify this, I must estimate the counterfactual outcome of what would have happened to the subfield had it not been included in a CPG. In particular, one threat to identification is root articles and their associated subfields may ask more important questions or use more rigorous methodology, resulting in them being more likely to be included in the CPG, and these characteristics might also be related to the observed results. To address this challenge, I match potential control articles on a carefully selected set of article characteristics prior to inclusion in the guideline. The ideal controls should satisfy three criteria: 1) it is unaffected by inclusion of the treated root articles in the CPG, 2) it has similar characteristics to treated articles before inclusion in the CPG, and 3) it must exhibit a parallel trajectory to what the treated subfields would have had were they not been included in the CPG. I employ a difference-in-differences estimator to study the within-subfield change in follow-on research before and after inclusion in a CPG. This approach relies on the assumption the timing of CPG inclusion is exogenous: that there was no other event, precisely at the time of guideline inclusion, which influenced both guideline inclusion and subsequent subfield research.

In particular, one might be worried if my controls are topically closely related to treated articles, they might be negatively impacted by their exclusion from the CPG. Were this true, this would violate the stable unit treatment values assumption (SUTVA) and lead to positively biased estimates of treatment effect. It is reassuring this is not a large concern that only 1.8% of CV disease publications are cited by a CPG. On the other hand, if my controls are distantly related, they might not show a parallel research trajectory to treated subfields.

For these reasons, my preferred control group is based upon matching to coarsened CV disease topics. I capitalize on the hierarchical structure of MeSH to identify all 425 MeSH terms under the MeSH term *cardiovascular disease*. These range from the common CV diseases (e.g. *myocardial infarction* or *atrial fibrillation*) to the more esoteric (e.g. *Lutembacher syndrome* or *Takotsubo cardiomyopathy*). I map the 75 CV disease terms at the third level of the MeSH hierarchical tree to one of 16 coarsened CV disease categories (see details in Appendix C). This approach balances sufficient granularity, as the largest category represents 17.5% of all CV papers while the smallest is 1.1%, with practicality. Since papers may be characterized by multiple CV MeSH terms and thus be members of multiple disease categories, matching against indicator variables for each of these 75 MeSH terms would create 3.8×10^{22} possible combinations of disease MeSH

⁸ This norm is also present in clinical medicine, but may not be as strong (Baerlocher et al. 2007, Fong and Wilhite 2017). For this reason, I repeated author-level analysis focusing on solely the first author or the means across all authors with similar results obtained.

(many of which would be empty sets). This approach takes seriously the idea the boundary between diseases is not always a bright line, but also accounts for the natural history of some diseases causing them to be closely interrelated. A heart attack may be caused when a clot blocks blood flow in a coronary artery. A closely related disease process, with its own distinct MeSH term, is myocardial reperfusion injury, which can occur when the clot is removed; this approach considers these to be within the same disease category. In contrast, a common cause of heart failure is high blood pressure. High blood pressure is a common medical condition, many of whom never develop heart failure, and yet the management of high blood pressure with heart failure has distinct considerations from that in an otherwise healthy individual. This approach would identify as separate those papers on solely heart failure, solely high blood pressure, and both heart failure and high blood pressure.⁹ I will later explore alternative controls group, including both those very distantly and very closely related to the treated articles.

More specifically, my preferred empirical strategy relies on selecting a matched control root article and associated subfield for each treated root article and subfield. To construct a plausible counterfactual of articles and their associated subfields which potentially could have been included in a guideline but were not, I employ coarsened exact matching (Blackwell et al. 2009, Iacus et al. 2011) (see Appendix C for details). I consider the set of potential controls to be the 831,218 original research articles indexed in PubMed 2011 or before with at least one CV disease MeSH term. I carefully match on a set of article characteristic prior to inclusion in a CPG including: 1) publication year, 2) cumulative citations as well as citations in the year prior to treatment to capture citation trajectory, 3) journal impact factor, 4) reference age as a measure of distance to the research frontier, 5) number of original research references for the extant of knowledge built upon, 6) a set of 16 indicator variables for coarsened CV disease topics, and 7) a set of 4 indicator variables for research methodology. For each article selected this way, I then identify its associated subfield. Given a potential control article could be matched at multiple different time periods—an article published in 1985 not included in a guideline in 1990 is also available as a control with an updated set of pre-treatment covariates in 1995—I performed one-to-one matching without replacement. Importantly, articles which are themselves later cited by a CPG are considered valid controls.¹⁰ Appendix Figure C2 provides an example of a matched treated and control root article.

Econometric considerations

⁹ Consider the 2005 ACC/AHA CPG on heart failure (Hunt et al. 2005). It cites 507 original research articles, 464 (91.5%) of which have at least one CV disease. Among these, 393 (84.5%) are associated with heart failure disease MeSH, but the next two most prevalent CV disease categories include 56 (12.1%) papers with arrhythmia MeSH and 36 (7.8%) with hypertension.

¹⁰ Three hundred seventy (4.5%) of controls selected by coarsened exact matching are later cited by a U.S., Canadian, U.K., or Australian medical professional society CPG written by 2014 in any field of medicine. Dropping observations after these controls are themselves treated by a CPG does not substantially alter the magnitudes of the observed treatment effect.

My primary estimating equation relates the dependent variable of interest, Y_{it} , of subfield i of CPG j in year t . Formally, I estimate variations of

$$E[Y_{it}|X_{it}] = \exp[\beta_1(Treated_i \times AfterCPG_t) + \beta_2 AfterCPG_t + f(AGE_{it}) + \delta_i + \gamma_t]$$

where $Treated_i$ is an indicator variable equal to 1 if a subfield was cited by a CPG and 0 otherwise. The term δ_i represents the individual fixed effects of subfields and controls for time-invariant subfield attributes, while the term γ_t represents a full set of calendar year fixed effects to account for changes in publication trends over time. I include a flexible function of subfield age, $f(AGE_{it})$, to control for trends in the evolution of subfields as they are born and mature.¹¹ My main coefficient of interest is β_1 , which captures the marginal effect of CPG inclusion and is identified by within-article variation in citation rates and between-article variation in the timing of CPG inclusion.

I also include an indicator variable, $AfterCPG_t$, equal to 1 in the years after a subfield’s root article is first cited by a CPG. As control subfields are uniquely linked to treated subfields through the matching process, they inherit a counterfactual date of “treatment” from their paired treated subfield. This, combined with the staggered treatment dates from CPGs, allows me to separately identify $AfterCPG_t$ from the main effect of interest, $Treated_i \times AfterCPG_t$. Inclusion of this term addresses the concern that age, year, and subfield fixed effects may not fully account for shifts in subfield activity around the time of inclusion in a CPG (Jaravel et al. 2018).

Since many of the outcomes studied are counts with highly skewed and non-negative distributions, following longstanding practice in the analysis of bibliometric data, I use a conditional fixed-effect Poisson model estimated by quasi-maximum likelihood (Hausman et al. 1984, Wooldridge 1997). For outcomes that are not counts, I estimated fixed effect linear regression. I use robust standard errors throughout, clustered at the level of the subfield.

Descriptive Statistics

Table 2 presents descriptive statistics and covariate balance of the matched treated and control root articles (Appendix Table A2 has these measures for the full populations of treated root articles and potential controls). A total of 8,275 (55.4%) of the 14,930 root articles under study were matched. By construction, treated and control papers are exactly matched on the publication year, treatment year, coarsened CV disease topic, and research methodology. While some residual difference remain, such as across the number of authors, references and articles within a subfield prior to treatment, it is reassuring the balance for

¹¹ Subfield age is calculated as the difference between the year of interest and the year a subfield is “born.” In my primary specification, I define birth as the year of the 5th publication in the subfield to reduce the impact of outlier publications. The findings are robust to using other thresholds, including using the year the first article in the subfield is published.

variables not used in the matching process, such as number of authors or references, is much improved (Appendix Table A2). The matching process tended to exclude outlier papers: while the mean pre-treatment citations for the full population of root articles was 50.1, it decreased to 15.3 in the subsample of matched root articles. One potential concern with matching is while the means may be similar across groups, the underlying distribution could be quite different. Figure 1 shows this is not the case, with treated and control root articles having broadly similar distributions of pre-treatment citations, reference mean ages, and subfield size.

The small differences in means of some pre-treatment characteristics could be problematic for my identification strategy if they reflected differing trends between treated and control root articles. I will show this is not the case using a regression framework below. Examination of the raw annual flow of articles into the associated subfields for treated and control root articles, particularly in the years surrounding treatment, suggests the parallel trends assumption is not violated (Appendix Figure A2).

Recall a possible concern to identification is my control subfields might themselves be impacted by treatment. Appendix Figure A3 plots the distribution in the number of shared articles between the associated subfields of treated root articles and their matched controls prior to treatment. In line with the idea my approach coarsely matches on topic, the mean fractional overlap between treated and control subfields is 1.1% of articles, with 7,263 (87.8%) of subfield treated and control pairs having zero overlap. My findings are robust to the exclusion of overlapped articles.

5. Results

Effect of CPG on subfield growth, attention, and scientific impact

I first aim to measure how CPGs alter the growth of the subfield. Table 3 reports the results from a difference-in-differences analysis for 8,275 treated and 8,275 control root articles. Treated subfields have a 9.5% increase in the flow of original research articles into the subfield following the CPG (column 1). Likewise, citations to *any* article within the subfield increase by 13.4% after the CPG (column 2), consistent with the subfield also receiving increased attention.

In Figure 2 I examine the dynamics of these effects. To do this, I estimate a specification in which the treatment effect is interacted with a set of indicator variables corresponding to year relative to the CPG, and then graph the resulting coefficients and ninety-five percent confidence intervals. Two features stand out. First, there is no discernible difference in trends in the years prior to treatment, lending support to my identification strategy. Second, the magnitude of the impact is sustained over the entire time period examined.

Having established an increase in research in a subfield, I next seek to understand the impact of this growth. Are the new papers in the subfield of marginal quality, or are they important contributions to

the field? To answer this, I characterize each article in a subfield by its vintage-specific long-run citations.¹² Separating the articles into 6 mutually exclusive bins based on their citation quantile, a striking pattern emerges (Figure 3).¹³ The treatment effect is fairly constant across the bottom three quartiles of the citation distribution. The bottom quartile, for instance, has an 8.0% increase in the flow of articles after the CPG. In contrast, articles above the ninety-ninth percentile in vintage-specific citations have a 42.2% increase in the flow of articles; the results are highly significant. Taken together, this shows that not only do subfields grow in size, but they become more impactful as measured by citations following the CPG.

CPGs and the direction of research within a subfield

I next focus on measuring the direction of innovation within a subfield. A subfield may grow in size and the attention it receives from researchers, but have the research be very similar in topic and methodology to what came before. Alternatively, the research may shift towards either a more incremental approach, doubling down on small improvements, or shift to a more exploratory path where researchers search for and test more intellectually distant ideas. I develop a set of measures to characterize the degree of exploration and exploitation.

As a first approach, I examine the level of evidence articles within a subfield provide. In medicine, there is a hierarchy in the level of evidence provided by different study methodologies (Burns et al. 2011). The highest quality evidence is provided by large, randomized controlled clinical trials, however, these are often time consuming, expensive, and require sufficient prior evidence to justify the equanimity in exposing patients to a clinical intervention. In contrast, a case report requires only a single patient and describes what happened in the course of routine care. In between these extremes, are various types of observational studies, include prospective cohorts and comprehensive retrospective case series. Translational and bench research may be considered pre-clinical and more investigational. If the effect of a CPG were a shift towards incremental research, one might expect researchers would respond by trying to provide increasingly higher levels of evidence to support the CPG recommendation. In other words, there would a shift away from case studies and towards clinical trials. Instead, I find the opposite: the lower the level of supporting evidence a research methodology provides, the larger the increase in articles within a subfield following a CPG (Table 4, panel A). For instance, clinical trials and epidemiologic studies do not have a statistically significant increase, although the former are imprecisely estimated since clinical trials are infrequent. In contrast, translational research has a 21.3% increase in the flow of articles and case reports a 10.8% increase.¹⁴

¹² Citation percentiles are computed relative to the universe of the over 17 million articles listed in both PubMed and Web of Science.

¹³ As this measure requires citation data from Web of Science, only those articles within a subfield that are published 2014 or before (when this data is available) and are indexed in Web of Science are included in the analysis.

¹⁴ Some research types remain infrequent even if they experience large increases. For instance, only 4.0% of all articles within treated and control subfields during the study period were translational.

As a complement to examining the methodology of the subfield article itself, I instead focus on the degree of evidence in a subfield articles references (Appendix Table A3, Panel A). For each paper, I calculate an evidence index, which is simply the number of clinical trial references minus the number of case report references, normalized by the total number of original research references. I find there is a greater increase in the flow of articles that draw on lower levels of evidence in their references.

I next move to relating a focal article and its original research references in idea space. First, an article may build on work broadly, recombining from many different research threads, or build on a set of highly similar, narrowly focused articles. To capture this, I parse articles as being above or below the median for the concentration of the associated MeSH terms of their references, calculated analogously to the Herfindahl-Hirschman index. I find a larger increase in articles that build on a more intellectually diverse set of reference articles (Table 4, Panel B, columns 1 and 2). Similarly, an article can have a high degree of intellectual overlap in the work it builds on, or it can be more intellectual distant. I capture this by calculating the average fraction of total MeSH shared between a focal article and its original research references, again parsed at the median value across the sample. I find a larger increase in those articles that are more intellectually distant from the references on which they build (Table 4, Panel B, columns 3 and 4).¹⁵

Research may either be consolidative, reinforcing earlier results, or disruptive, overturning the status quo. To provide evidence of this, I employ the disruption index proposed by Funk and Owen-Smith (2017). In this index, a highly disruptive article is one in which future research cites the focal paper but not the focal paper’s references, while highly consolidative research has future research citing both the focal paper and the focal paper’s references. Table A3, Panel B focuses on articles in the tail distribution of this index, those articles below the fifth percentile and above the ninety-fifth percentile. While there is no statistically significant change in articles in the bottom fifth percentile, there is a large increase in the flow of articles above the ninety-fifth percentile.¹⁶

While the previous measures demonstrated a greater increase in more exploratory research, I also try to measure distinctively incremental research. As one approach to this, researchers may take a finding validated in particular setting or population and test it in a different setting or population. For instance, prior work may have shown a particular drug is effective for middle-aged American men. This drug could then be tested in elderly men, children, or a middle-aged men in a different country. Such work is important for the advancement of clinical medicine, but is incremental in nature. To detect the signature of this type of research, I count those clinical studies (those tagged with disease but not bench science MeSH) which are the first time a particular population or geography MeSH keyword is used within the subfield, and compare

¹⁵ The findings with both of these measures are robust to exclusion of all basic science articles.

¹⁶ I explored several different variants of the disruption index, including altering the number of shared references required for a consolidative link and eliminating adjustments from papers that cite the focal article’s references but not the focal article itself, all with similar results obtained (Bornmann et al. 2020).

this to articles using the same set of keywords but not for the first time within the subfield (see Appendix C for details). There is a small increase in this distinctly incremental research (Table 4, Panel C).

As an alternative approach to measuring distinctly incremental research, I focus on work seeking to operationalize earlier research findings, moving them from upstream of patient care towards routine clinical use. To measure this, I construct a set of MeSH keywords measuring healthcare provider clinical skills, procedures, diagnostics, treatment, and care delivery (see Appendix C for details). Care must be taken on how any particular one of these measures is interpreted. For instance, a CPG may recommend that pharmacologic rather than surgical management is superior; the resulting decrease in papers tagged with surgical keywords should not be construed as moving further from the patient. For this reason, I focus on the overall measure of clinical studies tagged with at least one MeSH keyword from any of these five categories. I find there is a larger increase for clinical studies possessing none of these keywords than those possessing at least one (Appendix Table A3, Panel C).

Taken together these findings suggest that while there is a small increase in incremental research, the primary effect of CPGs is to shift the direction of research within a subfield towards exploration with work which is more translational, builds on less established prior research, draws on more diverse ideas, is more intellectual distant, and is more disruptive.

Commercial orientation of subfield research

Another way the direction of innovation can change is to have a greater commercial orientation. This is indeed the case following a CPG (Table 5). To capture this, I identify those articles cited by a patent at least once (Marx and Fuegi 2020). Columns 1 and 2 show that after the CPG, subfield articles are more likely to have been cited by a patent. I also examine author affiliations in Web of Science to identify those subfield articles having at least one author with an industry affiliation (see Appendix C for details). While there is a larger increase in the flow of articles with an industry-affiliated author compared to those with none, the difference between the two is not precisely estimated.

Which subfields are most impacted by CPGs?

As a final step in the analysis, I seek to understand which subfields are most responsive to knowledge certification. In particular, technological maturity may be an important margin. Established technologies, which have far less underlying dynamic and static uncertainty, may have less room for further technological improvement and may respond less to knowledge certification. Alternatively, technologies which are more immature—those which were only recently developed and have substantial opportunities for refinement—may be characterized by greater static and dynamic uncertainty and so may have a larger response to knowledge certification. Crucially, recall the implications of a shift towards exploration or exploitation are very different at different stages of a technology’s development, as a premature shift towards exploitation

may forestall the development of promising technological trajectories and decrease or delay ultimate technological performance (Gans et al. 2020).

To explore this I use three related measures of technological maturity. First, I measure the age of the subfield, defining a young field as under 15 years. Second, since clinical trials in patients require extensive prior research to justify and represent a more established subfield, I separate the subfields into those having at least one prior clinical trial and those having none at the time of the CPG. Finally, I examine the concentration of original research references among articles within the subfield prior to the CPG, again parsing the subfields at the median value. This measures the degree of coherence the subfield, where a more mature subfield might have greater internal coherence and agreement on the literature on which it draws, while an immature subfield might draw from more diverse prior work. Integrating across these measures, I find the consistent pattern that more technologically immature subfields experience a greater increase in the flow of articles into the subfield than those that are more mature (Table 6). For example, subfields with a prior clinical trial do not show an increase in articles following a CPG, while those without a prior trial have an 11.1% increase. I also explore how the degree of technological maturity of the subfield affects the extent to which the CPG stimulates incremental compared to exploratory research. Across the same measures for the direction of research used above, I find more technologically immature fields have a much larger increase in exploratory research than incremental research following a CPG. For example, there is a 4.7 vs. 1.3% increase in clinical studies using new population or geography keywords, but a 32.5% increase vs. 17.6% decrease for translational research following a CPG for subfields without a prior clinical trial and with a prior clinical trial, respectively.

Alternative mechanisms

While I will argue the primary effect of CPGs is mediated through knowledge certification altering the dynamic and static uncertainty of building on the knowledge, CPGs may shape the amount and direction of innovation through other potential mechanisms.

As widely disseminated documents, CPGs may work through diffusion and simply enhanced visibility of the knowledge they contain. I take two approaches to examining this. First, I focus on the dynamics of who responds to CPGs (Table 7a). I focus on the last author's expertise within cardiology. I parse subfield articles by the percentile of the last author's cumulative citations to their earlier CV disease original research articles; authors are considered new to cardiology research if they have never previously published such an article. Two patterns emerge. First, the largest increase is from "star" researchers within cardiology, those with the highest number of cumulative citations to CV disease publications. Second, there is also a significant increase in authors new to cardiology research.¹⁷ To expand on this, I parse subfield articles by whether the last author has previously published within the subfield. While there is a significant

¹⁷ I find a similar pattern when measuring author expertise using citations to all prior original research articles instead of limiting to just CV articles (Table A4a).

increase from authors new to the subfield, the increase in the flow of papers into the subfield is much larger for those authors who have previously published in the field (Table A4b). This suggests rather than authors with limited prior knowledge about the subfield, it is precisely those who would be arguably best informed about the subfield—those who have previously published in it and star cardiology researchers—who have the greatest increase in publications within the subfield following the CPG.¹⁸

As a second approach to the role of attention in mediating the results, I examine response heterogeneity by root article prominence (Table 7b). I separate the sample based upon the vintage-adjusted citation percentile of the root article in the year prior to treatment by the CPG; root articles cited by a CPG in the same year they are published are excluded. I find no clear pattern between prior citations of the root article and the flow of articles into the subfield. In particular, those subfields whose root article was above the 95th percentile in vintage-adjusted citations prior to the CPG—those root articles that were already highly prominent—experienced a treatment effect very similar to the overall sample. This suggests the subfield results are not mediated by an increase in attention. I supplement this by repeating the analysis, but using citations to the root article as the dependent variable. I find clear evidence for increased attention in mediating the increase in citations to root articles following a CPG. This speaks directly to how the choice to cite a root article and author a paper in the same subfield have different incentives and constraints.

Another mechanism through which the effect of CPGs on innovation could be mediated is codification. By codify knowledge, CPGs reduce the access cost for researchers who would otherwise have to spend substantially more time to sift through and evaluate the literature themselves. To empirically test this, I identify the 11,135 (67%) root articles cited by a CV review article prior to their inclusion in the CPG. While the prominence and resulting attention from a CV review article and professional society CPG are very different, these articles all previously received codification. Table 8 shows there is no difference in the treatment effect for articles codified prior to inclusion in a CPG and those that were not.

Robustness analysis

I also perform a number of robustness checks to assess the sensitivity of my estimates to alternative assumptions and subsamples. Appendix Table A5 explores a “dose effect,” as measured by the number of distinct CPGs citing a particular root article. While 6,837 (69.7%) of matched treated root articles are cited by a single CPG, 1,784 (18.2%) are cited by two CPGS, 749 (7.6%) by three, and 440 (4.5%) by four or more. Across a range of outcomes, stronger treatment effects are observed the greater the number of CPGs citing a root article. While this result must be treated with caution, since there is positive selection by CPGs for which articles are included multiple times, this supports the observed treatment effect being due to CPG inclusion and not some other shock sharing similar timing to CPG inclusion.

¹⁸ These findings are robust to exclusion of CPG authors.

Appendix Table A6 explores two different placebo tests. First, I randomly assign a mock treatment year during the pre-treatment period to root articles, subject to the constraint of at least 3 years of data available to establish trends in both the pre-treatment and post-treatment periods (column 1 and 2). I find no effect, supporting the parallel trends assumption during the pre-treatment period. As a second placebo test, I examine the 3,926 matched treated root articles that were dropped from a CPG, defined as an article included in a CPG but not cited by the next updated CPG in the guideline series. Using the year the root article is dropped as the year of treatment, I find no effect relative to control articles (columns 3 and 4). This both supports the parallel trends assumption and argues against a shock occurring at the same time as CPG inclusion affecting the trajectory of follow-on research.

As an additional test of my empirical strategy, including the parallel trends assumption, rather than focusing on the flow of articles into the subfield, I examine how inclusion in the CPG shapes citations to the root article (Appendix Figure A4 and Table A7). Similar to Figure 2, there is no discernible difference in trends in the years prior to treatment, lending support to my identification strategy. I also find CPGs act to diffuse the underlying knowledge root articles embody with a greater number of researchers citing this work and their follow-on research visible to a wider audience by virtue of its publication in a greater number of journals.

One threat to identification is the potential presence of a preceding change in a field leading both to receiving a CPG and the observed changes to innovation in the field. Recall it typically takes 1-2 years from when a decision is made to write a particular CPG and its publication. If there was such a shock leading to both the CPG and change in innovation, given this delay, it should be clearly visible in the pre-treatment period. No such change, however, is discernable in the raw time trends and treatment effect dynamics (Figures 2, A2, and A4). Additionally, since my controls are coarsely matched on disease topic, any such shock would apply to them as well. To further explore this, I plotted the raw number of total publications (Figure A5) and patents (Figure A6) for the disease topics of the first four CPGs in my sample. There is no clear increase in the number of publications and patents in the years prior to the CPG, arguing against the presence of such a shock.

Table A8 explores alternative control groups (see Appendix C for details on their construction and Table A2 for descriptive statistics). One potential concern is articles not included in a CPG may be negatively impacted by their exclusion. Were this true, this would violate the stable unit treatment values assumption (SUTVA) and lead to positively biased estimates of treatment effect. Recall in my preferred control group, treated and control root articles were matched on 16 coarsened CV disease categories. Fifty-six percent of controls share no CV disease MeSH with their treated root article. The results are robust to repeating the analysis limiting it to just these articles with no shared CV disease MeSH (row 2). As a second approach, I instead use pulmonary disease articles as a control, matched on the relative frequency of the MeSH within all CV and pulmonary disease publications for treated and control articles, respectively. There should be little concern these control articles are negatively impacted by their exclusion from CV

CPGs. The estimated impact on subfield article flow is smaller in magnitude and significant at the $p < 0.10$, however, remains positive (row 4). In particular, however, the dynamics observed in Figure A5 strongly suggest this control approach violates the parallel trends assumption. For instance, from 1980 to 2010, there is over a 1300% increase in the annual number of articles about heart failure, but only a 339% increase in articles about myocardial infarction.

On the other side, if the base case controls are not closely related enough to the treated articles, they may violate the parallel trends assumption. To address this, I limit the analysis to the 27% of treated and control pairs sharing the exact same CV disease MeSH (row 3). As an alternative approach, I construct an entirely new control sample requiring an exact match between the most specific CV disease MeSH term, as measured by its frequency among all CV disease publications (row 5). In both cases, I obtain similar results.

Finally, CEM by construction drops those observations not sharing a common support with potential control articles. Because of this, CEM estimates the treatment effect for the matched root articles, which may depart from the average treatment effect in the full sample due to these dropped observations. To address this potential concern, I repeated the analysis using nearest neighbor original research articles in the same journal issue as the root articles for controls. These articles are not balanced on pre-treatment observables (Table A2), however, they have the advantage of allowing me to estimate a treatment effect for the full sample of 14,930 root articles rather than the 8,275 matched to controls via CEM. Across a range of outcomes variables, broadly similar treatment effects are observed (row 6).

Appendix A includes a number of other robustness checks, including: exclusion of articles citing the CPG or subfield articles citing the root article (Table A9), choice of functional form (Table A10), two-period OLS (Table A11), and analysis at the level of the CPG rather than subfield (Table A12). As heterogeneous treatment effects may lead to bias in two-way fixed effect difference-in-differences estimators, I also include a “stacked regression” which is robust to these concerns (Table A13) (Baker et al. 2022, Goodman-Bacon 2021).

6. Discussion

In this paper, I leverage the rich bibliometrics of medical innovation and the tools of causal inference to explore the role of knowledge certification by CPGs—intended to improve the quality and decrease the cost of clinical care delivery—on innovation. I find CPGs had a large and unintended impact on innovation. In addition to diffusing the underlying knowledge they include, CPGs lead to a large and sustained increase research activity within included subfields. The work within subfields is not only more translational, draws from more diverse and less established prior research, and is more intellectual distant, but it is also more scientifically impactful. In particular, more technologically immature subfields show a greater shift in the amount and direction of innovation following a CPG than more established subfields. This suggests

knowledge certification can serve as an anchor for future innovation. Rather than leading primarily to more incremental work, it can provide a solid foundation from which to explore, allowing follow-on researchers to take more risks.

It is striking I did not find evidence knowledge certification decreased any particular type of innovation, but instead increased both more incremental and exploratory research. This is consistent with the idea that exploration and exploitation are not always strongly dichotomous processes (Sarnecka and Pisano 2020). My findings also stand in contrast to authors studying closely related phenomena such as codification (Garcia-Muina et al. 2009) and standards (Farrell and Saloner 1985) who worry such processes may preferentially support incremental innovation. In contrast to concerns shifts towards incremental innovation early in a technology’s course might be detrimental by curtailing the exploration of promising development pathways, I instead find the opposite: the more immature the technology, the greater the relative shift towards exploratory research. At first blush, it may seem surprising knowledge certification has a greater effect on breakthrough innovation. This is consistent, however, with prior research showing radical innovation is based to a higher degree rather than lower degree on existing knowledge (Schoenmakers and Duysters 2010, Uzzi et al. 2013) and local search (Jung and Lee 2015). It also may reflect that in some settings knowledge certification leads to a greater reduction in uncertainty for technological immature subfields and high-risk research trajectories than for established technologies or more incremental research trajectories.

Studying the direction of technological change poses significant measurement challenges. A strength of this study is the granularity and transparency MeSH-based measures of research topic and direction provide. Yet, measures of intellectual distance are necessarily coarse. While I find there is more exploration, I can say less about the size of the intellectual space in which this exploration is carried out and to what extent this exploration is itself constrained.

It can be informative to contrast CPGs with standards. Standards, especially where there are technological interoperability concerns, have received substantial attention from economists in how they impact innovation (David and Greenstein 1990, Farrell and Saloner 1985, Rysman and Simcoe 2008, Wen et al. 2015). Standards may also be established for the delivery of services, where they are implemented to improve the quality, decrease the cost, or decrease unwanted variability (Blind 2004). While CPGs might be considered a standard for the delivery of a service, unlike how standards are typically conceived, physician have substantial bounded autonomy in both the care of patients and research to depart from the CPGs (Mello 2001). These findings may not generalize to settings where knowledge certification is accompanied by sharp constraints on subsequent innovative activity.

It is difficult to compare the magnitude of the effect of CPGs on innovation directly with their effect on clinical care. The studies of CPG impact on clinical care occur across a range of practice settings and diseases and frequently feature a direct intervention to push providers towards compliance with CPG recommendations (Grimshaw et al. 1995, Lugtenberg 2009). It is not infrequent, however, for studies to

report 10-20% improvements in clinical care quality measures after active interventions to increase compliance with CPG recommendations. This places the impact of CPGs on innovation as roughly of the same order of magnitude as their impact on clinical care. Similar to Finkelstein’s study of policies to increase usage of preexisting vaccines (2004), the dynamic effects of CPGs on innovation are large and unintended.

What mechanisms mediate the observed treatment effect from CPGs? It is likely that multiple different mechanisms play a role, however, my empirics can shed light on their relative importance. CPGs do codify knowledge, acting to decrease the transaction cost of accessing it. Previously codified articles having a similar treatment effect as those that have not been codified argues strongly this is not a first-order process in this setting. Additionally, the evidence around the shift in research direction seems difficult to reconcile if the primary mechanism was simply a decrease in transaction costs. Similarly, CPGs certainly act to increase the visibility of the work they include, diffusing the underlying knowledge. However, the largest effect size from CPGs was observed from authors who had previously published within the same subfield and star cardiologists, precisely those who would already be best informed about this knowledge. In addition, the lack of a strong connection between the effect of certification and prior attention a root article has received suggests this is also not a first-order process. A further potential mechanism is that CPGs may act as a focusing device: by altering how healthcare providers treat patients, CPGs may change which problems they encounter in the daily practice of medicine and hence what they see as worth studying (Rosenberg 1969). I do not observe the clinical activity of researchers to be able to test this directly. The treatment effect from CPGs, however, is visible in the first year. In contrast, medicine is known to have a very slow diffusion of changes to clinical practice, typically measured on the order of decades, arguing against this as a driving mechanism (Morris et al. 2011, Skinner and Staiger 2007). Additionally, prior medical literature has highlighted that passive dissemination of information to change healthcare provider behavior is generally ineffective, and providers are frequently non-compliant with CPG recommendations (Grimshaw et al. 2001, Milchak et al. 2004). If this mechanism is prominent, I would expect the impact on innovation to be smaller than the impact on clinical care, as on the margin only some fraction of providers who change their clinical care due to the CPG would then go on to innovate in response to it. This is not what I observe. Finally, social influence and professional society authority could potentially play a role. However, CPGs had a larger effect on bench research and technologically immature subfields, locations where I might expect these forces to be less prominent than for clinical research and mature subfields, respectively. I argue the primary mechanism at work is a reduction in the static and dynamic uncertainty of building on certified knowledge. The empirical signature of static and dynamic uncertainty is similar in this context, and this study cannot distinguish their relative role.

Knowledge certification should not be viewed as a single phenomenon. Certification efforts may vary in the rigor of the process to evaluate and synthesize knowledge, the authority and status of the certifying body, and other such dimensions. It remains an open question the relative contribution from these different aspects to the observed treatment effect. In addition to professional medical society CPGs,

other settings have similarly exacting processes of synthesis and evaluation for the knowledge they certify. This includes the United Nations Climate Change Report, architectural standards, and manufacturing practices such as Total Quality Management. It is unclear how much dilution in the strength of the certification processes is allowed before the treatment effect goes away.

While this study focused on CPGs and biomedicine, the challenge of static and dynamic uncertainty is a fundamental problem for innovation across different domains. A similar dynamic may occur for businesses selecting technological innovation projects. Analogous to individual scientists selecting research projects, firms and managers of innovation are concerned with the expected returns, risks of failed projects, and reputational consequences when selecting which uncertain technological innovation projects to pursue. Perhaps more in the realm of supposition, business model innovations also have significant static and dynamic uncertainty and may face similar challenges (Amit and Zott 2012, Chesbrough 2009). Efforts to certify knowledge and de-risk the accompanying fundamental assumptions may help support more exploratory business model innovations. In this respect, this study speaks to the value of knowledge certification as a component of a firm’s knowledge management strategy (Hansen et al. 1999, Nonaka and Hirotaka 1995).

More generally, one view of innovation emphasizes the value of experimentation, “failing fast,” and the option value of potential innovative pathways (Camuffo et al. 2020, Reis 2011). This study shows how knowledge certification—that can require significant investments in human capital, resources, and time to complete—facilitates innovation that is more radical. These investments are very much a form of commitment. In line with work on the importance of long-term horizons and tolerance for failure (Azoulay et al. 2011, Bernstein 2015, Tian and Wang 2014), this suggests in some settings radical innovation may in particular benefit from commitment rather than experimentation (Gans et al. 2020).

From a policy perspective, like knowledge more generally, certified knowledge is non-rivalrous and partially non-excludable, with large positive externalities from its development (Foray 2004). To the extent knowledge certifiers are unable to fully internalize the value of their activity, knowledge certification will be under supplied by market mechanisms. This under provision is exacerbated by the long time frames and high degree of uncertainty frequently involved in innovative efforts making it difficult to accurately value efforts to certify knowledge *ex ante* (Rosenberg 1995). Policy and institutions can help lead to the optimal amount, content, and quality of knowledge certification.

The amount of knowledge innovators must master to reach the technological frontier is ever increasing (Jones 2009). As both the stock of existing knowledge and flow of new knowledge increases each day, this suggests the search and transaction cost of identifying and evaluating relevant knowledge is increasing. Knowledge certification may provide a partial solution to this challenge, allowing innovators to rely on the judgment of certifiers to more quickly reach the technological frontier rather than needing to synthesize and evaluate this knowledge themselves.

Finally, this study found a differential impact on innovation based on the degree of technological maturity. This suggests an often under-appreciated aspect in the study of technological change is the characteristics of the technology being studied itself. Teodoridis et al. (2019) found the role of generalist and specialists differed based on the pace of technological change, and Puranam et al. (2006) argued the effect of organizational structure after acquisition of small technology-based firms depended on the developmental stage of the acquired technology. My results suggest similar shocks can produce very different impacts on future technological development depending on the type of technology, its associated risks, and state of technological maturity. I expect further efforts to decompose how technology characteristics influence the impact of institutions and organization environment on innovation to be a fruitful line of future research.

References

- Amit, Raphael, and Christoph Zott. 2012. "Creating Value Through Business Model Innovation." *MIT Sloan Management Review*, 53(3):41-49.
- Azoulay, Pierre, Christian Fons-Rosen, and Joshua S. Graff Zivin. 2019. "Does Science Advance One Funeral at a Time?" *American Economic Review*, 109(8):2889-2920.
- Azoulay, Pierre, Jeffrey L. Furman, Joshua L. Krieger, and Fiona Murray. 2015. "Retractions." *Review of Economics and Statistics*, 97(5):1118-1136.
- Azoulay, Pierre, Joshua S. Graff Zivin, and Gustavo Manso. 2011. "Incentives and creativity: evidence from the academic life sciences." *RAND Journal of Economics*, 42(3):527-554.
- Azoulay, Pierre, Joshua Graff Zivin and Jialan Wang. 2010. "Superstar extinction." *Quarterly Journal of Economics*, 125(2):549-589.
- Azoulay, Pierre, Wesley H. Greenblatt, and Misty L. Heggeness. 2021. "Long-term Effects from Early Exposure to Research: Evidence from the NIH "Yellow Berets." *Research Policy*, 50(9):104332.
- Bachrach, C.A., and Thelma Charen. 1978. "Selection of Medline contents, the development of its thesaurus, and the indexing process." *Medical Informatics*, 3(3):237-254.
- Baerlocher, Mark Otto, Marshall Newton, Tina Gautam, George Tomlinson, and Allan Detsky. 2007. "The meaning of author order in medical research." *Journal of Investigative Medicine*, 55(4):174-180.
- Baker, Andrew C., David F. Larcker, and Charles C. Y. Wang. 2022. "How much should we trust staggered difference-in-differences estimates?" *Journal of Financial Economics*, 144(2):370-395.
- Balconi, Margherita. 2002. "Tacitness, codification of technological knowledge and the organization of industry." *Research Policy*, 31(3):357-379.
- Begley, C. Glenn, and John P. A. Ioannidis. 2015. "Reproducibility in science: improving the standard for basic and preclinical research." *Circulation Research*, 116(1):116-126.
- Benner, Mary J., and Michael L. Tushman. 2003. "Exploitation, Exploration, and Process Management: The Productivity Dilemma Revisited." *Academy of Management Review*, 28(2):238-256.
- Bernstein, Shai. 2015. "Does going public affect innovation?" *Journal of Finance*, 70(4):1365-1403.
- Bikard, Michael, and Isabel Fernandez-Mateo. 2022. "Standing on the Shoulders (Male) Giants: Gender Inequality and the Technological Impact of Scientific Ideas." Working paper.
- Blackwell, Matthew, Stefano Iacus, Gary King, and Giuseppe Porro. 2009. "CEM: Coarsened Exact Matching in Stata." *The Stata Journal*, 9(4):524-546.
- Blind, Knut. 2004. *The Economics of Standards: Theory, Evidence, Policy*. Northampton, MA:Edward Elgar.
- Bornmann, Lutz, Sitaram Devarakonda, Alexander Tekles, and George Chacko. 2020. "Are disruption index indicators convergently valid? The comparison of several indicator variants with assessments by peers." *Quantitative Science Studies*, 1(3):1242-1259.
- Brook, Robert H. 1989. "Practice Guidelines and Practicing Medicine. Are they Compatible?" *Journal of the American Medical Association*, 262(21):3027-3030.
- Burns, Patricia B., Rob J. Rohrich, and Kevin C. Chung. 2011. "The Levels of Evidence and their role in Evidence-Based Medicine." *Plastic and Reconstructive Surgery*, 128(1):305-310.
- Butler, Declan. 2008. "Translational Research: Crossing the Valley of Death." *Nature*, 453(7197):840-842.
- Camuffo, Arnaldo, Alessandro Cordova, Alfonso Gambardella, and Chiara Spina. 2020. "A Scientific Approach to Entrepreneurial Decision Making: Evidence from a Randomized Control Trial." *Management Science*, 66(2):564-586.
- Cerfolio, Robert J. 2018. "Innovation can emerge from a culture of standardization." *Journal of Thoracic and Cardiovascular Surgery*, 156(2): 819.
- Chassin, Mark R., Jacqueline Kosecoff, R.E. Park, Constance M. Winslow, Katherine L. Kahn, and Nancy J. Merrick, et al. 1987. "Does inappropriate use explain geographic variations in the use of health care services? A study of three procedures." *Journal of the American Medical Association*, 258(18):2533-2537.
- Chesbrough, Henry. 2009. "Business Model Innovation: Opportunities and Barriers." *Long Range Planning*, 43(2-3):354-363.
- Christensen, Clayton M. 1997. *The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail*. Boston, MA: Harvard Business School Press.

- Cohen, Wesley M., and Daniel A. Levinthal. 1990. "Absorptive Capacity: A New Perspective on Learning and Innovation." *Administrative Science Quarterly*, 35(1):128-152.
- Coletti, Margaret H. and Howard L. Bleich. 2001. "Medical Subject Headings Used to Search the Biomedical Literature." *Journal of the American Medical Informatics Association*, 8(4):317-323.
- Council of Medical Specialty Societies. 1987. "Standards of Quality in Patient Care: The Importance and Risks of Standard Setting." In: Proceedings of an Invitational Conference. Washington, D.C., September 25-26, 1987.
- Dance, Amber. 2012. "Who's on First?" *Nature*, 489(7417): 591-593.
- Dasgupta, Partha, and Paul A. David. 1994. "Toward a new economics of science." *Research Policy*, 23:487-521.
- David, Paul A. and Shane Greenstein. 1990. "The Economics of Compatibility Standards: An Introduction to Recent Research." *Economics of Innovation and New Technology*, 1:3-41.
- Ebben, Remco HA, Lilian CM Vloet, Michael HJ Verhofstad, Sanne Meijer, Joke AJ Mintjes-de Groot, and Theo van Achterberg. 2013. "Adherence to guidelines and protocols in the prehospital and emergency care setting: a systematic review." *Scandinavian Journal of Trauma, Resuscitation, and Emergency Medicine*, 21:9-21.
- Eddy, David M., and John Billings. 1988. The quality of medical evidence: implications for quality of care. *Health Affairs*, 7(1):19-32.
- Ehls, Daniel, Karim R. Lakhani, and Jacqueline N. Lane. 2021. "A Field Experiment on the Role of Constraints in Creative Problem-Solving." *Academy of Management Proceedings*, 2021(1).
- Farrell, Joseph and Garth Saloner. 1985. "Standardization, compatibility, and innovation." *RAND Journal of Economics*, 16(1):70-83.
- Finkelstein, Amy. 2004. "Static and Dynamic Effects of Health Policy: Evidence from the Vaccine Industry." *The Quarterly Journal of Economics*, 119(2):57-564.
- Flodgren, Gerd, Amanda M Hall, Lucy Goulding, Martin P Eccles, Jeremy M. Grimshaw, Gillian C. Leng, and Sasha Shepperd. 2016. "Tools developed and disseminated by guideline producers to promote the uptake of their guidelines." *Cochrane Database of Systematic Reviews*, 22(8):CD010669.
- Fong, Eric A., and Allen W. Wilhite. 2017. "Authorship and citation manipulation in academic research." *PLoS ONE*, 12(12):e0187394.
- Foray, Dominique. 2004. *The Economics of Knowledge*. Cambridge, MA: MIT Press.
- Foster, Jacob G., Andrey Rzhetsky, and James A. Evans. 2015. "Tradition and Innovation in Scientists' Research Strategies." *American Sociological Review*, 80(5):875-908.
- Franzoni, Chiara, and Paula Stephan. 2021. "Uncertainty and Risk-Taking in Science: Meaning, Measurement, and Management." NBER Working Paper #28562.
- Funk-Brentano, Christian. 2006. "Beta-blockade in CHF: from contraindication to indication." *European Heart Journal*, 8(Supplement C):C19-C27.
- Funk, Russell J., and Jason Owen-Smith. 2017. "A Dynamic Network Measure of Technological Change." *Management Science*, 63(3):791-817.
- Funk, Mark E., and Carolyn Anne Reid. 1983. "Indexing consistency in MEDLINE." *Bulletin of the Medical Library Association*, 71(2):176-183.
- Furman, Jeffrey L. and Scott Stern. 2011. "Climbing atop the Shoulders of Giants: The Impact of Institutions on Cumulative Research." *American Economic Review*, 101(5):1933-1963.
- Gans, Joshua S. 2016. *The Disruption Dilemma*. Cambridge, MA: MIT Press.
- Gans, Joshua S., Michael Kearney, Erin L. Scott, and Scott Stern. 2020. "Choosing Technology: An Entrepreneurial Strategy Approach." *Strategy Science*, 6(1):1-16.
- Garcia-Muina, Fernando E., Eva Pelechano-Barahona, and Jose Emilio Navas-Lopez. 2009. "Knowledge codification and technological innovation success: Empirical evidence from Spanish biotech companies." *Technological Forecasting and Social Change*, 76:141-153.
- Gelijns, Annetine, and Nathan Rosenberg. 1994. "The Dynamics of Technological Change in Medicine." *Health Affairs*, 13(3):28-46.
- Goodman-Bacon, Andrew. 2021. "Differences-in-differences with variation in treatment timing." *Journal of Econometrics*, 225(2):254-277.
- Grimshaw, Jeremy, Nick Freemantle, Sheila Wallace, Ian Russell, Brian Hurwitz, and Ian Watt, et al. 1995. "Developing and Implementing Clinical Practice Guidelines." *Quality in Healthcare*, 4:55-64.

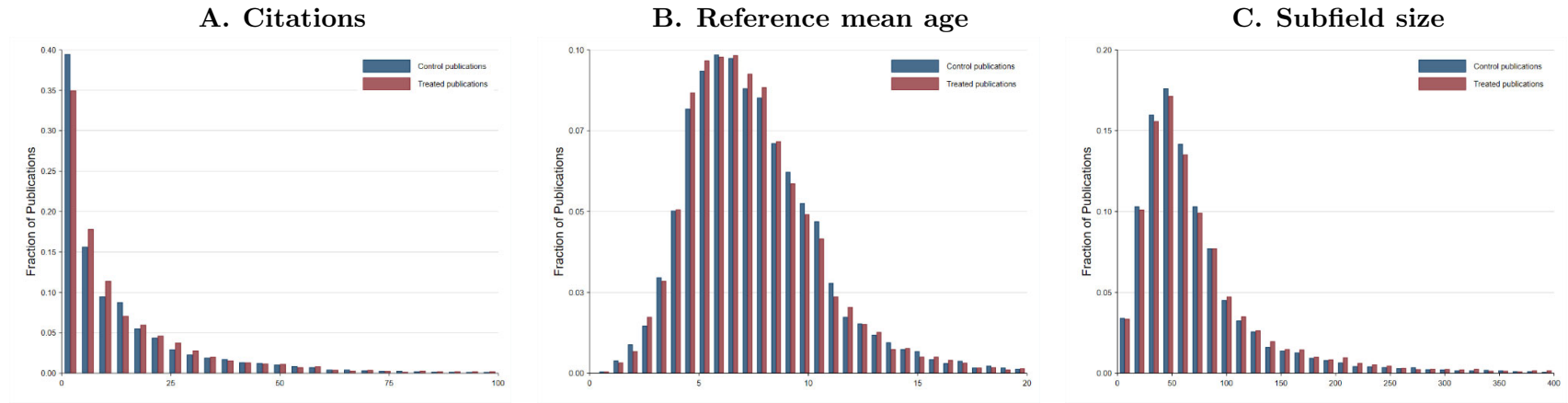
- Grimshaw, Jeremy M., Liz Shirran, Ruth Thomas, Graham Mowatt, Cynthia Fraser, and Lisa Bero, et al. 2001. "Changing Provider Behavior: An Overview of Systematic Reviews of Interventions." *Medical Care*, 39(8):II2-II45.
- Hansen, Morten T., Nitin Nohria, and Thomas Tierney. 1999. "What's Your Strategy for Managing Knowledge?" *Harvard Business Review*, 77(2):106-16.
- Hausman, Jerry, Bronwyn H. Hall, and Zvi Griliches. 1984. "Econometric Models for Count Data with an Application to the Patents-R&D Relationship." *Econometrica*, 52(4): 909-938.
- Heart Failure Society of America. 2000. "HFSA guidelines for management of patients with heart failure caused by left ventricular systolic dysfunction—pharmacological approaches." *Pharmacotherapy*, 20(5):495-522.
- Henderson, Rebecca M., and Kim B. Clark. 1990. "Architectural Innovation: The Reconfiguration of Existing Product Technologies and the Failure of Established Firms." *Administrative Science Quarterly*, 35(1):9-30.
- Hill, Ryan, and Carolyn Stein. 2021. "Race to the Bottom: Competition and Quality in Science." MIT Working Paper.
- Hughes, Robert A., Paul M. Gertman, Jennifer J. Anderson, Nurit L. Friedman, Mayde R. Rosen, Ann C. Ward, and Bernard E. Kreger. 1984. "The Ancillary Services Review Program in Massachusetts. Experience of the 1982 pilot project." *Journal of the American Medical Association*, 252(13):1727-32.
- Hunt, Sharon Ann, William T. Abraham, Marshall H. Chin, Arthur M. Feldman, Gary S. Francis, and Theodore G. Ganiats et al. 2005. "ACC/AHA 2005 Guideline Update for the Diagnosis and Management of Chronic Heart Failure in the Adult. A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Writing Committee to Update the 2001 Guidelines for the Evaluation and Management of Heart Failure)." *Circulation*, 112(12):e154-e235.
- Iacus, Stefano M., Gary King, and Giuseppe Porro. 2011. "Causal Inference without Balance Checking: Coarsened Exact Matching." *Political Analysis*, 20(1):1-24.
- Institute of Medicine. 1989. *Controlling Costs and Changing Patient Care? The Role of Utilization Management*. Washington, DC: The National Academies Press.
- Institute of Medicine. 1990. *Clinical Practice Guidelines: Directions for a New Program*. Washington, DC: The National Academies Press.
- Institute of Medicine. 1992. *Guidelines for Clinical Practice: From Development to Use*. Washington, DC: The National Academies Press.
- Institute of Medicine. 1995. *Setting Priorities for Clinical Practice Guidelines*. Washington, DC: The National Academies Press.
- Institute of Medicine. 2011. *Clinical Practice Guidelines We Can Trust*. Washington, DC: The National Academies Press.
- Ioannidis, John P. A. 2005. "Why Most Published Research Findings are False." *PLoS Medicine*, 2(8):e124.
- Jansen, Justin J. P., Frans A. J. Van Den Bosch, and Henk W. Volberda. 2006. Exploratory Innovation, Exploitative Innovation, and Performance: Effects of Organizational Antecedents and Environmental Moderators." *Management Science*, 52(11):1661-1674.
- Jaravel, Xavier, Neviana Petkova, and Alex Bell. 2018. "Team-Specific Capital and Innovation." *American Economic Review*, 108(4-5):1034-1073.
- Jones, Benjamin F. 2009. "The Burden of Knowledge and the "Death of the Renaissance Man": Is Innovation Getting Harder?" *The Review of Economic Studies*, 76(1):283-317.
- Jue, J. Jane, Sarah Cunningham, Kathleen Lohr, Paul Shekelle, Richard Shiffman, and Craig Robbins, et al. 2019. "Developing and Testing the Agency for Healthcare Research and Quality's National Guideline Clearinghouse Extent of Adherence to Trustworthy Standards (NEATS) Instrument." *Annals of Internal Medicine*, 170:480-487.
- Jung, Hyun Ju and Jeongsik "Jay" Lee. 2015. "The Quest for Originality: A New Typology of Knowledge Search and Breakthrough Inventions." *Academy of Management Journal*, 59(5):1725-1753.
- Koffi, Marlene, and Matt Marx. 2022. "Cassatts in the Attic." Working paper.
- Komajda, Michel, Pablo Lapuerta, Nancy Hermans, Jose Ramon Gonzale-Junatay, Dirk J van Veldhuisen, and Erland Erdmann, et al. 2005. "Adherence to guidelines is a predictor of outcome in chronic heart failure: the MAHLER survey." *European Heart Journal*, 26(16):1653-1659.

- Law, Yuk. 2017. "Does standardization improve care or stifle innovation?" *Pediatric Transplantation*, 21:e12773.
- Lee, Deishin, and Eric Van den Steen. 2010. "Managing Know-How." *Management Science*, 56(2):270-285.
- Lee, Dong Heun, and Ole Vielemeyer. 2011. "Analysis of Overall Level of Evidence Behind Infectious Diseases Society of America Practice Guidelines." *Archives of Internal Medicine*, 171(1):18-22.
- Lerchmueller, Marc J. and Olav Sorenson. 2016. "Author Disambiguation in PubMed: Evidence on the Precision and Recall of Author-ity among NIH-Funded Scientists." *PLoS One*, 11(7):e0158731.
- Lin, Jimmy, and W. John Wilbur. 2007. "PubMed Related Articles: A Probabilistic Topic-Based Model for Content Similarity." *BMC Bioinformatics*, 8: 423 (2007).
- Lissoni, Franceso. 2001. "Knowledge codification and the geography of innovation: the case of Brescia mechanical cluster." *Research Policy*, 20:1479-1500.
- Lugtenberg, Marjolein, J. S. Burgers, and G. P. Westert. 2009. "Effects of evidence-based clinical practice guidelines on quality of care: a systematic review." *Quality and Safety in Health Care*, 18(5):385-92.
- March, James G. 1991. "Exploration and Exploitation in Organizational Learning." *Organization Science*, 2(1):71-87.
- Marx, Matt, and Aaron Fuegi. 2020. "Reliance on science: Worldwide front-page patent citations to scientific articles." *Strategic Management Journal*, 41(9):1572-1594.
- McAlister, Finlay A., Sean van Diepen, Rajdeep S. Padwal, Jeffrey A. Johnson, Sumit R. Majumdar. 2007. "How Evidence-Based are the Recommendations in Evidence-Based Guidelines?" *PLoS Medicine*, 4(8):e250.
- McCann, Shaun R. 2020. "Can guidelines inhibit innovation and critical thinking?" *Bone Marrow Transplantation*, 55:1217-1219.
- McMahan, Peter, and Daniel A. McFarland. 2021. "Creative Destruction: The Structural Consequences of Scientific Curation." *American Sociological Review*, 86(2):341-376.
- Mechanick, Jeffrey I., Rachel Pessah-Pollack, Pauline Camacho, Ricardo Correa, M. Kathleen Figaro, and Jeffrey R. Garber, et al. 2017. "American Association of Clinical Endocrinologists and American College of Endocrinology Protocol for Standardized Production of Clinical Practice Guidelines, Algorithms, and Checklists-2017 Update." *Endocrinology Practice*, 23(8):1006-1021.
- Mello, Michelle M. 2001. "Of Swords and Shields: The Role of Clinical Practice Guidelines in Medical Malpractice Litigation." *University of Pennsylvania Law Review*, 149(3):645-710.
- Mihm, Jurgen, Christoph H. Loch, Dennis Wilkinson, and Bernardo A. Huberman. 2010. "Hierarchical Structure and Search in Complex Organizations." *Management Science*, 56(5):831-848.
- Milchak, Jessica L, Barry L. Carter, Paul A. James, and Gail Ardery. 2004. "Measuring Adherence to Practice Guidelines for the Management of Hypertension." *Hypertension*, 44(5):602-608.
- Mokyr, Joel. 2002. *The Gifts of Athena: Historical Origins of the Knowledge Economy*. Princeton, NJ: Princeton University Press.
- Morris, Zoe Slote, Steven Wooding, and Jonathan Grant. 2011. "The answer is 17 years, what is the question? Understanding time lags in translational research." *Journal of the Royal Society of Medicine*, 104(12):510-520.
- Murray, Fiona, Philippe Aghion, Mathias Dewatripont, Julian Kolev, and Scott Stern. 2016. "Of Mice and Academics: Examining the Effect of Openness on Innovation." *American Economic Journal: Economic Policy*, 8(1):212-252.
- Myers, Kyle. 2020. "The Elasticity of Science." *American Economic Journal: Applied Economics*, 12(4):103-134.
- Nelson, Richard. 1959. "The Simple Economics of Basic Science Research." *Journal of Political Economy*, 67(3):297-306.
- Nonaka, Ikujiro and Hirotaka Takeuchi. 1995. *The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation*. Oxford, UK: Oxford University Press.
- Pandyan, Anand D, Kate Radford, Stephen Ashford, Andrew Bateman, Christopher Burton, and Louise Connell, et al. 2013. "New guidelines on rehabilitation likely to restrict practices and stifle innovation." *British Medical Journal*, 347:f4876.
- Peltzman, Samuel. 1975. "The Effects of Automobile Safety Regulation." *Journal of Political Economy*, 83(4):677-726.
- Polanyi, Michael. 1966. *The Tacit Knowledge Dimension*. Chicago, IL:University of Chicago Press.

- Polidoro Jr., Francisco. 2013. "The Competitive Implications of Certifications: The Effects of Scientific and Regulatory Certifications on Entries into New Technical Fields." *Academy of Management Journal*, 56(2):597-627.
- Puranam, Phanish, Harbir Singh, and Murizio Zollo. 2006. "Organizing for Innovation: Managing the Coordination-Autonomy Dilemma in Technology Acquisitions." *Academy of Management Journal*, 49(2):263-280.
- Reis, Eric. 2011. *The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Business*. New York: Crown Business.
- Rosenberg, Nathan. 1969. "The Direction of Technological Change: Inducement Mechanisms and Focusing Devices." *Economic Development and Cultural Change*, 18(1):1-24.
- Rosenberg, Nathan. 1976. "On Technological Expectations." *The Economic Journal*, 86(343):523-535.
- Rosenberg, Nathan. 1995. "Uncertainty and Technological Change." *Federal Reserve Bank of Boston Conference Proceedings*, 40(Jun):91-125.
- Rosenthal, Tom. 2012. "Geographic variation in health care." *Annual Review of Medicine*, 63:493-509.
- Rosso, Brent D. 2014. "Creativity and Constraints: Exploring the Role of Constraints in the Creative Processes of Research and Development Teams." *Organization Studies*, 35(4):551-585.
- Rysman, Marc, and Timothy Simcoe. 2008. "Patents and the Performance of Voluntary Standard-Setting Organizations." *Management Science*, 54(11):1920-1934.
- Salandra, Rosella, Paola Criscuolo, and Ammon Salter. 2021. "Directing scientists away from potentially biased publications: the role of systematic reviews in health care." *Research Policy*, 50(1):104131.
- Sarnecka, Dominika K., and Gary P. Pisano. 2020. "The Evolutionary Nature of Breakthrough Innovation: Re-Evaluating the Exploration vs. Exploitation Dichotomy." Harvard Business School Working Paper 21-071.
- Sauermann, Henry, and Carolin Haeussler. 2017. "Authorship and contribution disclosures." *Science Advances*, 3(11): e1700404.
- Schoenmakers, Wilfred, and Geert Duysters. 2010. "The technological origins of radical inventions." *Research Policy*, 39(8):1051-1059.
- Scotchmer, Suzanne. 1991. "Standing on the Shoulders of Giants: Cumulative Research and the Patent Law." *Journal of Economic Perspectives*, 5(1): 29-41.
- Skinner, Jonathan, and Douglas Staiger. 2007. "Technology Adoption from Hybrid Corn to Beta-Blockers." In Berndt, Ernst R., and Charles R. Hulten (Eds). *Hard-to-Measure Goods and Services: Essays in Honor of Zvi Griliches*. Chicago, IL: University of Chicago Press.
- Sorenson, Olav, and Lee Fleming. 2004. "Science and the diffusion of knowledge." *Research Policy*, 33:1615-1634.
- Svorenčík, Andrej. 2014. "MIT's Rise to Prominence: Outline of a Collective Biography." *History of Political Economy*, 46(suppl 1): 109-133.
- Stephan, Paula. 2012. *How Economics Shapes Science*. Cambridge, MA: Harvard University Press.
- Stokes, Patricia D. 2005. *Creativity from Constraints: The Psychology of Breakthrough*. New York, NY: Springer Publishing Company.
- Stokes, Donald E. 1997. *Pasteur's Quadrant: Basic Science and Technological Innovation*. Washington, D.C.:Brookings Institution Press.
- Stuart, Toby E. and Joel M. Podolny. 1996. "Local Search and the Evolution of Technological Capabilities." *Strategic Management Journal*, 17(S1):21-38.
- Teodoridis, Florenta, Michael Bikard, and Keyvan Vakili. 2019. "Creativity at the Knowledge Frontier: The Impact of Specialization in Fast- and Slow-paced Domains." *Administrative Science Quarterly*, 64(4):894-927.
- Thompson, Neil C. and Douglas Hanley. 2017. "Science is Shaped by Wikipedia: Evidence from a Randomized Control Trial" Working paper.
- Tian, Xuan and Tracy Yue Wang. 2014. "Tolerance for Failure and Corporate Innovation." *Review of Financial Studies*, 27(1):211-255.
- 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. and Neil R. Smalheiser. 2021. Author-ity 2018 - PubMed author name disambiguated dataset. University of Illinois at Urbana-Champaign.

- Torvik, Vette I, Marc Weeber, Don R. Swanson, and Neil R. Smalheiser. 2003. "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.
- Tricoci, Pierluigi, Joseph M. Allen, Judith M. Kramer, Robert M. Califf, and Sidney C. Smith. 2009. "Scientific Evidence Underlying the ACC/AHA Clinical Practice Guidelines." *Journal of the American Medical Association*, 301(83):831-841.
- Tversky, Amos and Daniel Kahneman. 1973. "Availability: A heuristic for judging frequency and probability." *Cognitive Psychology*, 5(2):207-232.
- U.S. General Accounting Office. 1991. *Practice Guidelines: The Experience of Medical Specialty Societies*. GAO/PEMD-91-11. Washington, D.C.:General Accounting Office.
- Uzzi, Brian, Satyam Mukherjee Michael Stringer, and Ben Jones. 2013. "Atypical Combinations and Scientific Impact." *Science*, 342(6157):468-472.
- von Hippel, Eric. 1994. "'Sticky Information' and the Locus of Problem Solving: Implications for Innovation." *Management Science*, 40(4):429-439.
- Weisz, George, Alberto Cambrosio, Peter Keating, Loes Knaapen, Thomas Schlich, and Virginie J.Tournay. 2007. "The Emergence of Clinical Practice Guidelines." *The Milbank Quarterly*, 85(4):691-727.
- Wen, Wen, Christ Forman, and Sirkka Jarvenpaa. 2015. "How Do Open Standards Influence Inventive Activity? Evidence from the IETF" *Academy of Management Proceedings*, 1:15562.
- Wilke, Michael, R. F. Grube, and K. F. Brodmann. 2011. "Guideline-adherent initial intravenous antibiotic therapy for hospital-acquired/ventilator-associated pneumonia is clinically superior, saves lives and is cheaper than non guideline adherent therapy." *European Journal of Medical Research*, 16(7):315-323.
- Williams, Heidi L. 2013. "Intellectual Property Rights and Innovation: Evidence from the Human Genome." *Journal of Political Economy*, 121(1):1-27.
- Winslow, Constance M., David H Solomon, Mark R. Chassin, Jacqueline Kosecoff, Nancy J. Merrick, and Robert H. Brook. 1988. "The Appropriateness of Carotid Endarterectomy." *New England Journal of Medicine*, 318(12):721-727.
- Wockel, Achim, Christian Kurzeder, Verena Geye, Igor Novasphenny, Regine Wolters, and Manfred Wischnewsky, et al. 2010. "Effects of guideline adherence in primary breast cancer-a 5-year multi-center cohort study of 3976 patients." *Breast*, 19(2):120-7.
- Wooldridge, Jeffrey M. 1997. "Quasi-Likelihood Methods for Count Data." In M. Hashem Pesaran, and Peter Schmidt (Eds.), *Handbook of Applied Econometrics*, pp. 352-406. Oxford: Blackwell.
- Woolf, Steven H. 1990. "Practice Guidelines: A New Reality in Medicine." *Archives of Internal Medicine*, 150(9):1811-1818.
- Woolf, Steven H. 1992. "Practice guidelines, a new reality in medicine. II. Methods of developing guidelines." *Archives of Internal Medicine*, 152(5):946-52.
- Woolf, Steven H. 2008. "The Meaning of Translational Research and Why it Matters." *Journal of the American Medical Association*, 299(2): 211-213.

Figure 1. Distribution of pre-treatment root article characteristics

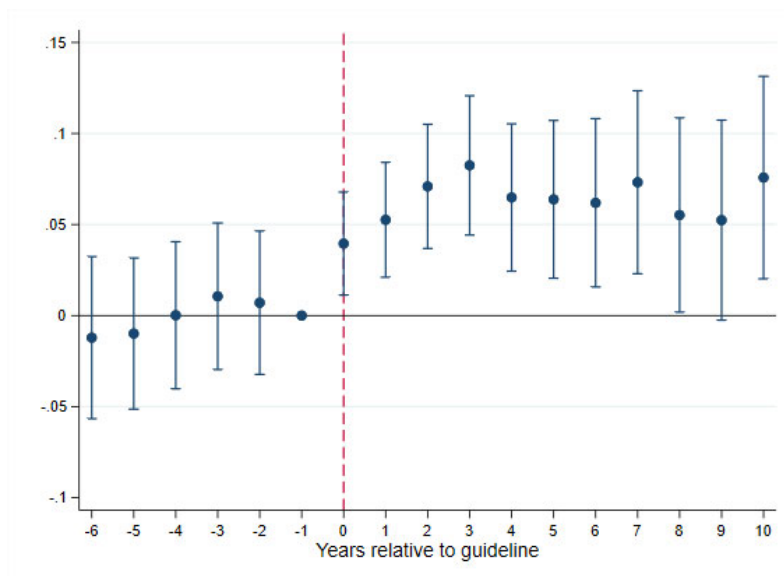


Note: Histogram of the number of citations from original research articles to each root article prior to treatment. Two hundred forty-six (123 treated, 123 controls) outlier root articles with over 100 citations are omitted.

Note: Histogram of the mean age, in years, of original articles referenced by each root article. Three hundred thirty-four with no references (167 treated, 167 controls) and 22 outliers with mean reference age over twenty years (11 treated, 11 controls) are omitted.

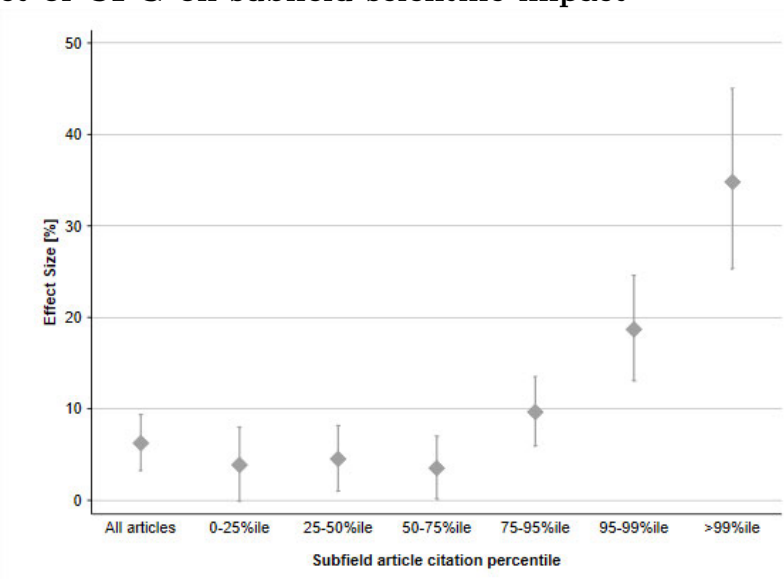
Note: Histogram of the number of original research articles in the subfield of each root article prior to treatment. One hundred seventy-eight (107 treated, 71 controls) outlier root articles with over 400 subfield articles are omitted.

Figure 2: Dynamic effects of CPG on subfield growth



Note: The plots above present results from a conditional fixed effects QML Poisson specification in which number of original research articles in a subfield is regressed onto 20 interaction terms between treatment status (having been cited by a guideline) and the number of years relative to treatment as well as subfield fixed effects and a set of 9 subfield age indicators. The year prior to treatment is omitted. Non-exponentiated coefficients are presented, with the vertical bars denoting the 95% confidence interval corresponding to robust standard errors, clustered at each subfield.

Figure 3. Effect of CPG on subfield scientific impact



Note: Each estimate corresponds to a separate conditional quasi-maximum likelihood Poisson regression model very similar to Table 3, column 1. The dependent variable is the flow of original research articles in the subfield by their vintage-adjusted citation quantile. Percentage increases, and their 95% confidence intervals computed on the basis of robust standard errors clustered at the subfield, are presented.

Table 1a. Descriptive statistics: Guidelines

	Mean	Median	Std. Dev.	Min.	Max.
Year	2004.430	2006	5.971	1990	2011
Author count	20.030	20	13.191	0	63
No individual authors	0.044	0	0.207	0	1
Citations	347.947	202	385.083	2	1,793
Number of references	243.022	124	280.526	1	1,194
Reference mean age (years)	7.129	7	2.943	1	16
Length (pages)	44.837	23	48.320	3	260
Receives or provides update to another guideline	0.748	1	0.436	0	1
Number of publishing journals	1.793	2	0.970	1	6
Number of sponsoring organizations	2.607	2	2.296	1	14
Sponsoring organization nationality					
Australia/New Zealand	0.067	0	0.250	0	1
Canada	0.378	0	0.487	0	1
United Kingdom	0.044	0	0.207	0	1
United States	0.600	1	0.492	0	1
Multinational sponsoring organizations	0.067	0	0.250	0	1

Note: N = 135 guidelines.

Table 1b. Descriptive statistics: Guideline series

	Mean	Median	Std. Dev.	Min.	Max.
Number of guidelines in a series	2.521	2	2.593	1	17
Ever received a guideline update	0.534	1	0.502	0	1
Average time between guideline updates	5.676	5	2.751	1	12

Note: N = 73 guideline series. Average time between updates is calculated conditional upon ever receiving an update (N=39).

Table 2. Pre-treatment characteristics of treated and control root articles

	Treated articles	Control articles	t-stat
Year published	1997.226 (0.087)	1997.226 (0.087)	0.000
Year treated	2004.141 (0.061)	2004.141 (0.061)	0.000
Citations prior to treatment	15.274 (0.283)	14.680 (0.291)	1.462
Citations in year preceding treatment	2.708 (0.048)	2.609 (0.050)	1.422
Journal impact factor	6.487 (0.072)	6.418 (0.070)	0.683
Author count	6.587 (0.037)	6.063 (0.035)	10.338
References count	18.444 (0.102)	18.826 (0.100)	2.674
Reference mean age (years)	7.419 (0.033)	7.483 (0.034)	1.368
Research methodology: Clinical trial	0.001 (0.000)	0.001 (0.000)	0.000
Research methodology: Case report	0.032 (0.002)	0.032 (0.002)	0.000
Research methodology: Epidemiologic study	0.450 (0.005)	0.450 (0.005)	0.000
Research methodology: Bench science	0.015 (0.001)	0.015 (0.001)	0.000
Coarsened CV Topic: Arrhythmias	0.148 (0.004)	0.148 (0.004)	0.000
Coarsened CV Topic: Valvular heart disease	0.062 (0.003)	0.062 (0.003)	0.000
Coarsened CV Topic: Hypertension	0.062 (0.003)	0.062 (0.003)	0.000
Year of first article in subfield	1966.775 (0.335)	1966.568 (0.338)	0.434
Number of articles in subfield prior to treatment	80.809 (1.035)	75.015 (0.841)	4.345

Note: N = 8,275 treated and 8,275 control articles selected after coarsened exact matching. Means, standard errors, and t-statistics are reported. Citations and references only include original research articles. Arrhythmias, valvular heart disease, and hypertension presented as representative cardiology disease topics.

Table 3. Effect of CPG on subfield growth and citations

	Articles	Citations
Treated $\times$ AfterCPG	1.095** (0.019)	1.134** (0.017)
Log pseudolikelihood	-808,788	-4,343,337
Number of root papers	16,526	16,549
Observations	333,442	305,265

Note: Estimates stem from conditional quasi-maximum likelihood Poisson regressions. The dependent variables are the number of original research articles in the subfield (column 1) and the number of citations from original research articles to original research article in the subfield (column 2). All models incorporate a full suite of subfield and year fixed effects as well as a series of 9 indicator variables for subfield age. Exponentiated coefficients are presented, which after subtracting one yield magnitudes interpretable as elasticities. Robust errors in parentheses, clustered at the subfield ($^{\dagger}p < 0.10$, $^*p < 0.05$, $^{**}p < 0.01$).

Table 4. Effect of CPG on subfield research direction

A. Research methodology						
	All articles	Clinical trial	Epidemiologic	Case reports	Translational	Bench science
Treated $\times$ AfterCPG	1.080** (0.017)	0.941 (0.074)	1.026 (0.019)	1.108** (0.036)	1.213** (0.049)	1.249** (0.051)
Log pseudolikelihood	-692,266	-15,827	-433,473	-105,194	-73,029	-84,833
Number of root papers	16,514	3,085	15,728	9,426	7,901	8,331
Observations	290,077	54,149	276,527	165,661	138,139	145,768

B. Intellectual diversity and distance				
	Reference MeSH concentration		MeSH overlap with references	
	Above median	Below median	Above median	Below median
Treated $\times$ AfterCPG	1.041* (0.019)	1.197** (0.021)	1.050** (0.017)	1.118** (0.021)
Log pseudolikelihood	-433,081	-540,255	-534,753	-455,399
Number of root papers	16,316	16,423	16,428	16,433
Observations	301,440	302,966	303,246	303,172

C. Incremental research topics						
	Populations		Geographies		Either	
	Old	New	Old	New	Old	New
Treated $\times$ AfterCPG	1.040* (0.018)	0.989 (0.023)	1.134** (0.040)	1.062** (0.022)	1.060 [†] (0.032)	1.028 [†] (0.015)
Log pseudolikelihood	-590,845	-74,024	-116,189	-164,890	-256,600	-522,699
Number of root papers	15,908	13,701	8,630	13,599	14,258	16,193
Observations	279,609	241,621	149,781	239,223	250,998	284,536

Note: Estimates stem from conditional quasi-maximum likelihood Poisson regressions. In all cases the dependent variable is the number of original research articles within the subfield by A. research methodology derived from MeSH keywords, B. the concentration of MeSH among a paper’s references and the mean fractional MeSH overlap between a paper and its references, and C. research topic derived from MeSH keywords (see section 5 and Appendix C for details of these measures). All models incorporate a full suite of subfield fixed effects as well as a set of 12 indicator variables for year and 5 indicator variables for subfield age. Exponentiated coefficients are presented, which after subtracting one yield magnitudes interpretable as elasticities. Robust errors in parentheses, clustered at the subfield ([†] $p < 0.10$, * $p < 0.05$, ** $p < 0.01$).

Table 5. Effect of CPG on subfield commercial orientation

	Ever cited by patent		Industry author affiliation	
	No	Yes	No author	At least one author
Treated $\times$ AfterCPG	1.071** (0.019)	1.200** (0.031)	1.108** (0.018)	1.155** (0.044)
Log pseudolikelihood	-781,918	-146,036	-645,725	-97,030
Number of root papers	16,524	13,667	16,515	10,748
Observations	333,406	275,988	304,683	195,869

Note: Estimates stem from conditional quasi-maximum likelihood Poisson regressions. The dependent variables are the number of original research articles within the subfield by if they have ever been cited by a patent (column 1 and 2) or have any author with an industry affiliation (columns 3 and 4). All models incorporate a full suite of subfield and year fixed effects as well as a series of 9 indicator variables for subfield age. Exponentiated coefficients are presented, which after subtracting one yield magnitudes interpretable as elasticities. Robust errors in parentheses, clustered at the subfield ($^{\dagger}p < 0.10$, $^*p < 0.05$, $^{**}p < 0.01$).

Table 6. Effect of CPG by subfield maturity

	Subfield age		Prior clinical trials		Reference concentration	
	Older	Younger	Any	None	More concentrated	Less concentrated
Nb. subfield articles	1.062** (0.023)	1.122** (0.025)	1.022 (0.050)	1.111** (0.017)	1.010 (0.022)	1.086** (0.025)
Research methodology						
Epidemiologic	0.986 (0.025)	1.101** (0.028)	0.995 (0.053)	1.045* (0.019)	1.031 (0.026)	1.011 (0.027)
Translational	1.208** (0.060)	1.227** (0.078)	0.824† (0.088)	1.325** (0.058)	0.842** (0.056)	1.172** (0.055)
MeSH overlap with references						
Above median	1.006 (0.022)	1.084** (0.027)	0.996 (0.047)	1.054** (0.018)	0.995 (0.022)	1.025 (0.025)
Below median	1.113** (0.028)	1.120** (0.030)	1.047 (0.062)	1.146** (0.021)	1.036 (0.028)	1.131** (0.028)
Incremental research topics						
Any new topics	1.011 (0.019)	1.061** (0.023)	1.013 (0.040)	1.047** (0.016)	1.004 (0.020)	1.025 (0.021)

Note: Each cell contains an estimate for the treatment effect in a separate regression. All estimates stem from conditional quasi-maximum likelihood Poisson regressions. In all cases the dependent variable is the number of original research articles within the subfield, characterized by measures identical to those in Table 4. The sample is split by those subfields under 15 years of age (column 1), 15 years and older (column 2), prior clinical trial in the subfield (column 3), no prior clinical trial in the subfield (column 4), and above (column 5) and below (column 6) median for the concentration of original research article references within the subfield prior to treatment. All models incorporate a full suite of subfield fixed effects as well as a set of 12 indicator variables for year and 5 indicator variables for subfield age. Exponentiated coefficients are presented, which after subtracting one yield magnitudes interpretable as elasticities. Robust errors in parentheses, clustered at the subfield († $p < 0.10$, * $p < 0.05$, ** $p < 0.01$).

Table 7a. Effect of CPG by last author expertise within cardiology

	New authors	0-50%ile	50-75%ile	75-95%ile	95-99%ile	>99%ile
Treated $\times$ AfterCPG	1.065** (0.022)	1.032 (0.021)	1.017 (0.021)	1.078** (0.019)	1.130** (0.021)	1.140** (0.022)
Log pseudolikelihood	-160,331	-149,695	-158,066	-298,623	-316,707	-418,719
Number of root papers	15,260	15,161	15,371	16,330	16,279	16,163
Observations	282,047	279,516	283,327	301,217	300,304	298,261

Note: Estimates stem from conditional quasi-maximum likelihood Poisson regressions. The dependent variables are the number of original research articles in the subfield by the last author's cumulative citation percentile for prior original research CV disease articles (columns 2-6) or if they have no prior CV disease original research articles (column 1). All models incorporate a full suite of subfield and year fixed effects as well as a series of 9 indicator variables for subfield age. Exponentiated coefficients are presented, which after subtracting one yield magnitudes interpretable as elasticities. Robust errors in parentheses, clustered at the subfield ($^{\dagger}p < 0.10$, $^*p < 0.05$, $^{**}p < 0.01$).

Table 7b. Effect of CPG by root article prominence

	All root articles	<25%ile	25-50%ile	50-75%ile	75-95%ile	>95%ile
Subfield article count	1.083** (0.018)	1.071 [†] (0.045)	1.023 (0.044)	1.096* (0.044)	1.084** (0.027)	1.062 [†] (0.036)
Root article citations	1.122** (0.024)	1.596** (0.170)	1.345** (0.080)	1.207** (0.040)	1.233** (0.033)	1.101** (0.041)

Note: Each cell contains an estimate for the treatment effect in a separate regression. All estimates stem from conditional quasi-maximum likelihood Poisson regressions. The dependent variable is the number of original research articles in the subfield (row 1) and number of citations from original research articles to the root article (row 2). The columns split the sample by the vintage-adjusted citation percentile of the root article in the year prior to treatment; root articles published in the same year they received treatment are excluded. All models incorporate a full suite of root article or subfield fixed effects as well as a set of 12 indicator variables for year and 5 indicator variables for age. Exponentiated coefficients are presented, which after subtracting one yield magnitudes interpretable as elasticities. Robust errors in parentheses, clustered at the subfield ($^{\dagger}p < 0.10$, $^*p < 0.05$, $^{**}p < 0.01$).

Table 8. Effect of CPG by previous codification in review articles

	Cited by cardiovascular review article prior to inclusion in CPG	
	No	Yes
Treated $\times$ AfterCPG	1.120** (0.035)	1.116** (0.023)
Log pseudolikelihood	-263,430	-542,656
Number of root papers	5,407	11,119
Observations	109,558	223,884

Note: Estimates stem from conditional quasi-maximum likelihood Poisson regressions. The dependent variables are the number of original research articles in the subfield. Column 1 estimates this for the subfields whose associated root article was not cited by a CV review article prior to treatment (2,115 treated subfields and 3,300 control subfields), while column 2 includes only those subfields where the associated root article was cited by a CV review article prior to treatment (6,160 treated subfields and 4,975 control subfields). All models incorporate a full suite of subfield and year fixed effects as well as a series of 9 indicator variables for subfield age. Exponentiated coefficients are presented, which after subtracting one yield magnitudes interpretable as elasticities. Robust errors in parentheses, clustered at the subfield ($^{\dagger}p < 0.10$, $^*p < 0.05$, $^{**}p < 0.01$).

Appendix A: Ancillary Results and Robustness Checks

Effect of CPG inclusion on the diffusion of knowledge

Examining citations offers a direct way of assessing diffusion of the knowledge included in a CPG, and is also a test of the validity of the empirical strategy. Table A7a reports these results. Overall, there is a 14.5% increase in citations from original research articles to root articles following inclusion in a CPG. This increase could be due to more intensive citation by authors who have previously built on the root article, or it could represent an increase in the extensive margin. To examine this, I look at citations by author and journal based on whether are new, meaning it is the first time a particular author or journal has cited the root article (columns 2 through 5). CPG inclusion leads to an increase in citations from new authors and journals similar in magnitude to the increase in those having previously cited the root paper, suggesting an increase in both the intensive and extensive margins. Finally, I calculate author and journal diversity following the index proposed by Teodoridis et al. (2019): 1 minus the Herfindahl index over all authors or journals each year. Both authors and journals have an increase in this measure of diversity (Table A7b).

In Figure A4, I examine the dynamics of these effects. To do this, I estimate a specification in which the treatment effect is interacted with a set of indicator variables corresponding to year relative to CPG inclusion, and then graph the resulting coefficients and ninety-five percent confidence intervals. Two features stand out. First, there is no discernible difference in trends in the years prior to treatment, lending support to my identification strategy. Second, the magnitude of the impact increases over time, suggesting the impact is sustained rather than transitory.

Together, these findings suggest CPG inclusion acts to help diffuse the underlying knowledge the root article embodies. A greater number of researchers build on the knowledge, and their follow-on research is itself visible to a wider audience by virtue of its publication in a greater number of journals.

Additional ancillary analysis and robustness checks

Some authors may seek to signal their quality by including citations from the CPG or altering how they position and write their work relative to articles included in the CPG. After exclusion of citations to root articles which themselves cite the treating CPG (columns 1 and 2 of Appendix Table A9) or subfield articles which also cite the root article (columns 3 and 4), the treatment is diminished in magnitude but still statistically significant.

In Appendix Table A10, I examine the sensitivity of my results to the choice of functional forms. I estimate a similar difference-in-differences model using fixed effect linear regression with outcomes transformed by the inverse hyperbolic sine function (Burbidge et al. 1988). Under this transformation, coefficient estimates can be interpreted as approximating elasticities. Across a range of outcomes, I find positive and precisely estimated effects of CPG inclusion.

Difference-in-differences estimation with panel data may be biased by serial correlation among the outcome variable across time (Bertrand et al. 2004). While quasi-maximum likelihood standard errors are robust to arbitrary patterns of serial correlation (Wooldridge 1997), I never the less estimate a two-period OLS model (Appendix Table A11). I find inclusion in a guideline leads to an annual increase of 0.732 citations to root articles and 0.241 articles within the subfield. At the mean values for these variables, this corresponds to a 34.8% increase in citations to root articles and a 4.7% increase in articles within the subfield.

Appendix Table A12 repeats the main result, but rolls up the data to the level of the CPG rather than subfield. With this approach, subfields for each CPG are lumped together, with controls inheriting a “counter-factual” guideline based on the CPG of their matched counterpart. This approach loses significant power given the small number of CPGs relative to subfields. While the result for the flow of articles is imprecisely estimated, it is similar in magnitude to analysis at the subfield level.

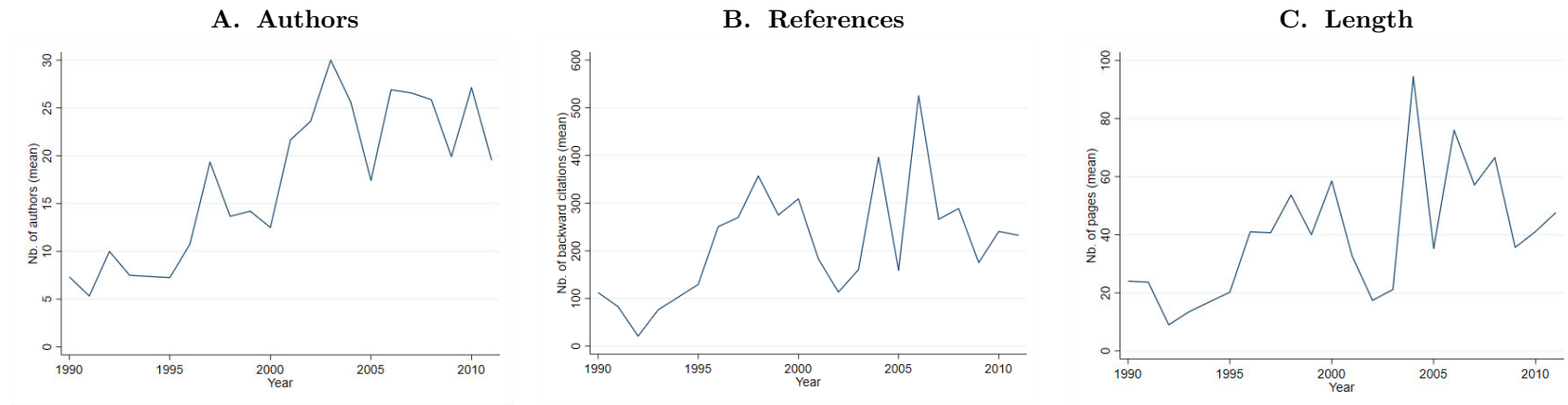
Recent literature has noted that when heterogeneous treatment effects are present two-way fixed effect difference-in-differences estimators may be biased (Baker et al. 2022, Goodman-Bacon 2021). A number of potential solutions have been proposed, including new estimators (Callaway and Sant’Anna 2021, de Chaisemartin and D’Haultfoeuille 2020, Sun and Abraham 2021), extended two-way (Mundlak) fixed-effects (Wooldridge 2021), and “stacked” regression (Cengiz et al. 2019, Deshpande and Li 2019). To address

this concern, I implement a stacked regression following Cengiz et al. (2019). For each guideline, I construct an event-specific “clean 2 x 2” where all controls are never treated during the observation window, only those subfields for which I observe 7 years before and 7 years after the guideline are included, and observations are aligned by event rather than calendar time. These individual clean 2 x 2 datasets are then stacked, with the resulting estimation similar to standard two-way fixed effect regression but includes dataset-specific unit- and time-fixed effects (Baker et al. 2022). Appendix Table A13 displays these results, showing my findings are robust to concerns of bias in two-way fixed effect estimators with heterogeneous treatment effects.

References

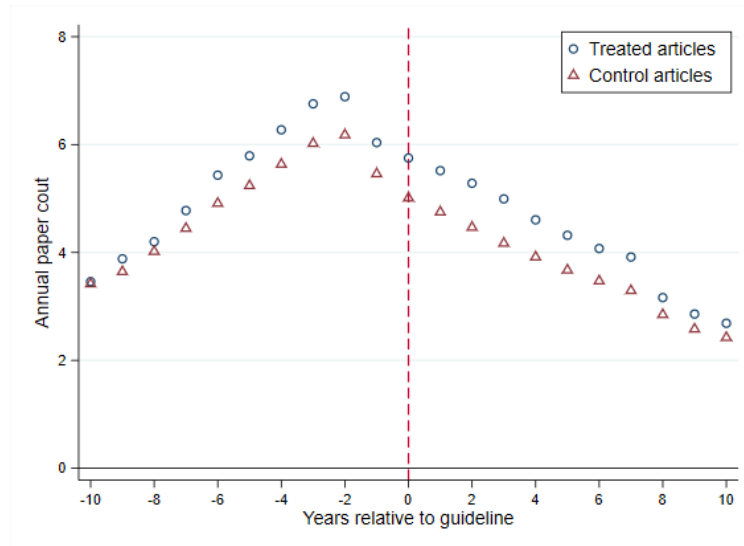
- Bertrand, Marianne, Esther Dufo, and Sendhil Mullainathan. 2004. “How Much Should We Trust Differences-in-Differences Estimates?” *Quarterly Journal of Economics*, 119(1):249–75.
- Burbidge, John B., Lonnie Magee, and A. Leslie Robb. 1988. “Alternative Transformations to Handle Extreme Values of the Dependent Variable.” *Journal of the American Statistical Association*, 83(401):123–27.
- Callaway, Brantly, and Pedro H.C. Sant’Anna. 2021. “Difference-in-Differences with multiple time periods.” *Journal of Econometrics*, 225(2):200–230.
- Cengiz, Doruk, Arindrajit Dube, Attila Lindner, and Ben Zipperer. 2019. “The Effect of Minimum Wages on Low-Wage Jobs.” *Quarterly Journal of Economics*, 134(3):1405–1454.
- De Chaisemartin, Clement, and Xavier D’Haultfoeuille. 2020. “Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects.” *American Economic Review*, 110(9):2964–2996.
- Deshpande, Manasi, and Yue Li. 2019. “Who is Screened Out? Application Costs and the Targeting of Disability Programs.” *American Economic Journal: Economic Policy*, 11(4):213–248.
- Sun, Liyang, and Sarah Abraham. 2021. “Estimating dynamic treatment effects in event studies with heterogeneous treatment effects.” *Journal of Econometrics*, 225(2):175–199.
- Wooldridge, Jeffrey M. 2021. “Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators.” Working paper.

Figure A1. Clinical practice guideline characteristics over time



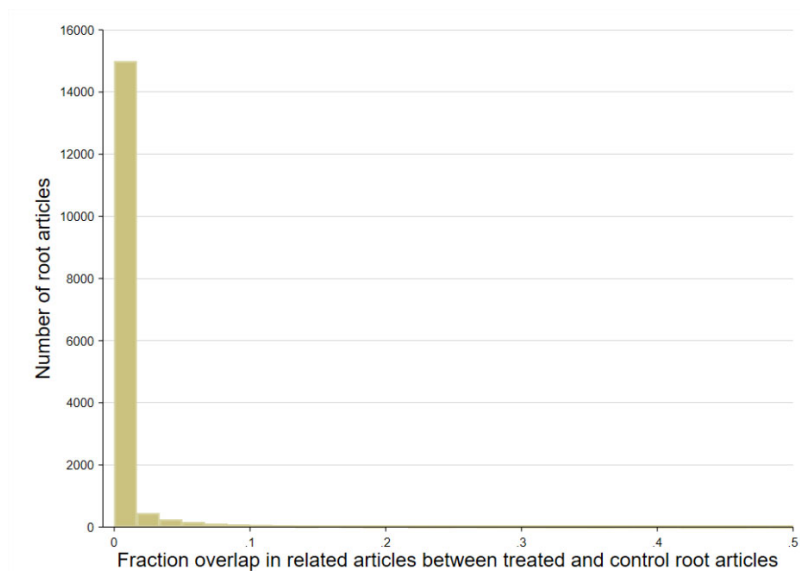
Note: N = 135 guidelines. Plots the mean (A) number of authors, (B) number of references, and (c) length in pages by year of clinical practice guideline publication.

Figure A2. Raw annual subfield article count over time



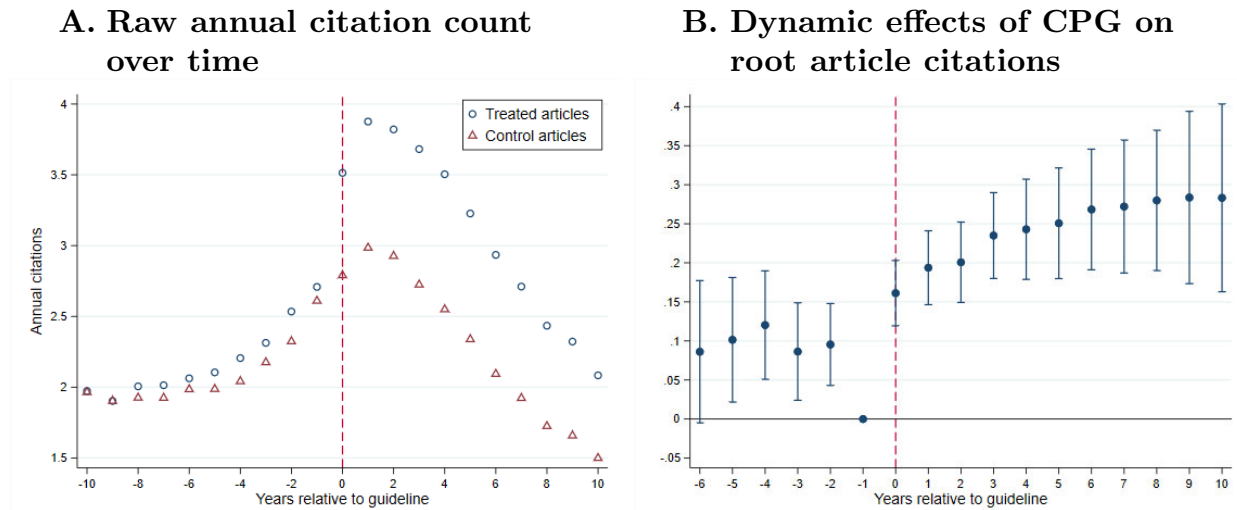
Note: The plot above presents the raw annual count of original research articles in the associated subfields for 8,275 treated root articles and 8,275 control root articles relative to the year of inclusion in a CPG.

Figure A3. Distribution of subfield overlap



Note: Histogram of the number of shared articles, prior to treatment, between a treated subfield and its control subfield selected via coarsened exact matching. Fifty outliers with a fractional overlap of over 0.5 were omitted. The mean fractional overlap is 1.1%, with 87.8% (14,526 of 16,550) having zero overlap between the subfields of the treated and control root articles.

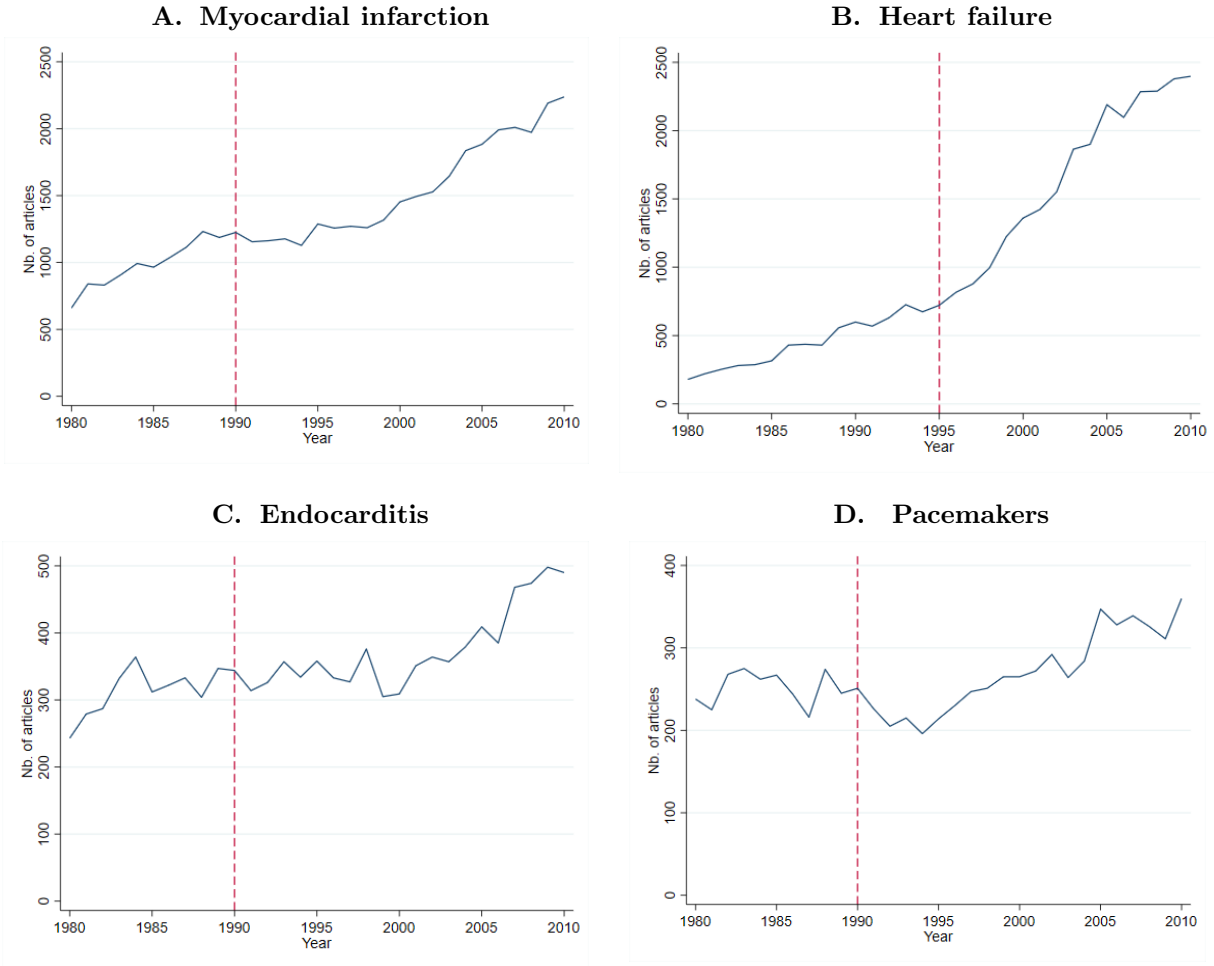
Figure A4. Effect of CPG on root article citations



Note: The plot above presents the raw annual citation count from original research articles to 8,275 treated root articles and 8,275 control root articles relative to the year of inclusion in a CPG

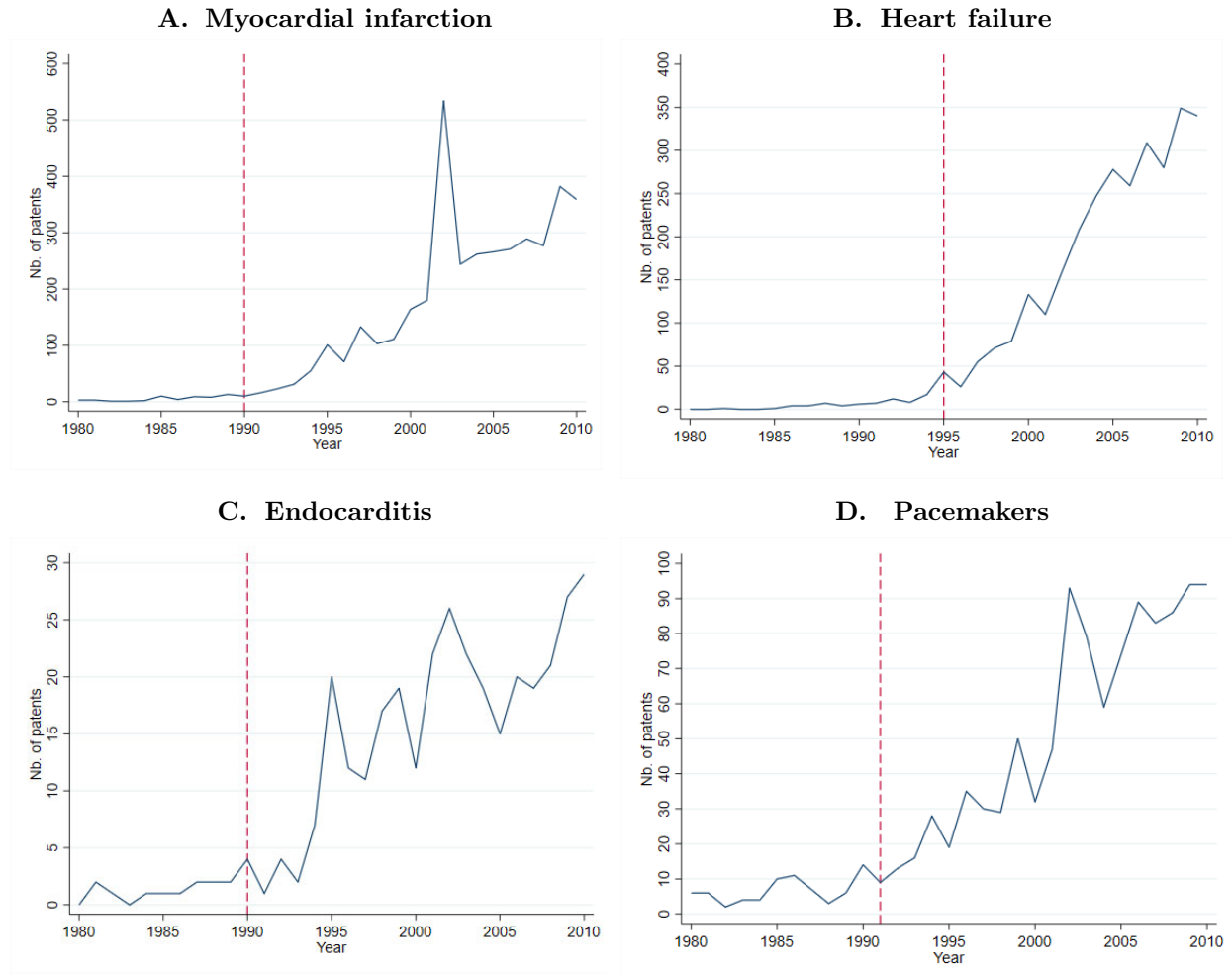
Note: The plots above present results from a conditional fixed effects QML Poisson specification in which citations from original research articles is regressed onto 20 interaction terms between treatment status (having been cited by a guideline) and the number of years relative to treatment as well as article fixed effects and a set of 9 root article age indicators. The year prior to treatment is omitted. Non-exponentiated coefficients are presented, with the vertical bars denoting the 95% confidence interval corresponding to robust standard errors, clustered at each root article.

Figure A5. Field trends over time: publication counts



Note: Plots the number of papers in PubMed, 1980-2010, with the corresponding term present in the title. The dashed vertical line represents the first year the topic received a CPG during the sample period.

Figure A6. Field trends over time: patent counts



Note: Plots the number of patents, by application year 1980-2010, which cite an article in PubMed with the corresponding term present in the title. The dashed vertical line represents the first year the topic received a CPG during the sample period.

Table A1. Descriptive statistics: Guidelines compared to review articles

	Guidelines		Review articles	
	Mean	Median	Mean	Median
Year	2004.430	2006	2002.528	2003
Journal impact factor	8.079	13	2.910	2
Author count	20.030	20	2.839	2
No individual authors	0.044	0	0.004	0
Citations	347.947	202	32.521	13
Number of references	243.022	124	52.108	40
Reference mean age (years)	7.129	7	7.683	7
Length (pages)	44.837	23	9.249	8

Note: Cardiology professional society clinical practice guidelines (N = 135) and review articles with at least one cardiovascular disease MeSH term (N = 138,537), 1990-2011.

Table A2. Covariate balance across different treatment and control samples

	All CV papers			CEM: Exact CV MeSH			CEM: Pulmonary controls			Nearest neighbor		
	Treated	Control	<i>t</i> -stat	Treated	Control	<i>t</i> -stat	Treated	Control	<i>t</i> -stat	Treated	Control	<i>t</i> -stat
Year published	1995.846 (0.074)	1993.973 (0.016)	16.005	1997.563 (0.099)	1997.563 (0.099)	0.000	1996.520 (0.085)	1996.520 (0.085)	0.000	1995.846 (0.074)	1995.779 (0.055)	0.729
Year treated	2003.964 (0.046)			2004.114 (0.070)	2004.114 (0.070)	0.000	2004.050 (0.055)	2004.050 (0.055)	0.000	2003.964 (0.046)	2004.040 (0.034)	1.329
Citations prior to treatment	29.545 (0.483)			12.762 (0.269)	12.035 (0.277)	1.884	15.795 (0.269)	15.064 (0.277)	1.895	29.545 (0.483)	15.008 (0.190)	33.073
Citations in year preceding treatment	4.719 (0.085)			2.354 (0.045)	2.224 (0.049)	1.957	2.596 (0.046)	2.477 (0.044)	1.877	4.719 (0.085)	2.161 (0.029)	34.674
Journal impact factor	8.885 (0.079)	3.778 (0.006)	116.338	5.890 (0.072)	5.831 (0.071)	0.581	6.286 (0.075)	6.517 (0.073)	2.205	8.885 (0.079)	8.677 (0.058)	2.119
Author count	6.564 (0.030)	4.922 (0.003)	64.559	6.558 (0.042)	6.047 (0.038)	9.092	6.321 (0.034)	5.683 (0.034)	13.405	6.564 (0.030)	5.740 (0.019)	24.117
Nb of references	18.970 (0.105)	15.499 (0.013)	34.870	18.462 (0.140)	18.775 (0.109)	1.768	17.991 (0.109)	18.097 (0.095)	0.737	18.970 (0.105)	17.334 (0.065)	13.954
Reference mean age (years)	7.593 (0.028)	8.390 (0.005)	21.930	7.362 (0.035)	7.449 (0.036)	1.740	7.714 (0.033)	7.776 (0.033)	1.326	7.593 (0.028)	7.788 (0.024)	5.133
Research method: Clinical trial	0.005 (0.001)	0.003 (0.000)	5.022	0.000 (0.000)	0.000 (0.000)	0.000	0.002 (0.000)	0.002 (0.000)	0.000	0.005 (0.001)	0.004 (0.000)	1.582
Research method: Case report	0.042 (0.002)	0.190 (0.000)	46.068	0.026 (0.002)	0.026 (0.002)	0.000	0.042 (0.002)	0.042 (0.002)	0.000	0.042 (0.002)	0.083 (0.002)	16.359
Research method: Epidemiologic study	0.479 (0.004)	0.208 (0.000)	80.578	0.448 (0.006)	0.448 (0.006)	0.000	0.438 (0.005)	0.438 (0.005)	0.000	0.479 (0.004)	0.314 (0.003)	34.090
Research method: Bench science	0.026 (0.001)	0.087 (0.000)	26.546	0.007 (0.001)	0.007 (0.001)	0.000	0.016 (0.001)	0.016 (0.001)	0.000	0.026 (0.001)	0.071 (0.002)	19.630
Year of first article in subfield	1964.616 (0.259)			1966.742 (0.379)	1966.235 (0.426)	0.891	1964.224 (0.316)	1960.154 (0.331)	8.900	1964.616 (0.259)	1962.841 (0.194)	5.463
Nb of subfield articles prior to tx	80.023 (0.725)			82.248 (1.171)	57.453 (0.894)	16.828	78.118 (0.879)	68.594 (0.734)	8.315	80.023 (0.725)	74.338 (0.476)	6.782
Number of articles	14,930	831,218		6,360	6,360		10,262	10,262		27,651	14,930	

Note: Means, standard errors, and *t*-statistics are reported. Citations, references, and subfield articles only include original research articles. Arrhythmias, valvular heart disease, and hypertension presented as representative cardiology disease topics. Columns 1-3 compare the full population of original research CV disease publications referenced by the 135 guidelines under study to all original research publications with a CV disease MeSH in PubMed. The following columns compare treated and control publications selected via coarsened exact matching requiring an exact match on the CV disease term (columns 4-6), coarsened exact matching using pulmonary disease articles as controls (columns 7-9), and nearest neighbors of treated articles (columns 10-12). CV—cardiovascular, CEM—coarsened exact matching, tx—treatment.

Table A3. Alternative measures of research direction

A. Degree of supporting evidence			
	Evidence index		
	Above median	Below median	
Treated $\times$ AfterCPG	1.063** (0.019)	1.114** (0.019)	
Log pseudolikelihood	-414,927	-571,609	
Number of root papers	16,439	16,449	
Observations	303,473	303,477	

B. Disruption index			
	<5%ile	5-95%ile	>95%ile
Treated $\times$ AfterCPG	0.949 (0.038)	1.072** (0.019)	1.148** (0.024)
Log pseudolikelihood	-57,704	-477,606	-427,873
Number of root papers	9,272	16,400	16,432
Observations	164,721	288,163	288,738

C. Distance from direct clinical care		
	MeSH of clinical study papers	
	Yes	No
Clinical skills	1.140** (0.048)	1.077** (0.019)
Procedures	1.127** (0.031)	1.074** (0.019)
Diagnostics	1.059** (0.019)	1.102** (0.025)
Treatment	1.055* (0.027)	1.111** (0.020)
Care delivery	1.024 (0.025)	1.052** (0.018)
Any	1.080** (0.020)	1.225** (0.029)

Note: Each cell contains an estimate for the treatment effect in a separate regression. All estimates stem from conditional quasi-maximum likelihood Poisson regressions. In all cases, the dependent variable is the number of original research articles within the subfield by A. evidence index, B. disruption index percentile, and C. research topic derived from MeSH keywords (see Appendix B for details of these measures). All models incorporate a full suite of subfield and year fixed effects as well as a series of 9 indicator variables for subfield age. Exponentiated coefficients are presented, which after subtracting one yield magnitudes interpretable as elasticities. Robust errors in parentheses, clustered at the subfield ($^{\dagger}p < 0.10$, $^*p < 0.05$, $^{**}p < 0.01$).

Table A4a. Effect of CPG by last author expertise

	New authors	0-50%ile	50-75%ile	75-95%ile	95-99%ile	>99%ile
Treated $\times$ AfterCPG	1.039 [†] (0.024)	1.038 (0.024)	0.980 (0.021)	1.038* (0.018)	1.114** (0.020)	1.153** (0.022)
Log pseudolikelihood	-110,651	-117,569	-128,020	-311,134	-381,731	-423,683
Number of root papers	13,823	14,111	14,583	16,354	16,395	16,377
Observations	256,244	260,602	269,182	301,701	302,411	302,216

Note: Estimates stem from conditional quasi-maximum likelihood Poisson regressions. The dependent variables are the number of original research articles in the subfield by the last author's cumulative citation percentile for all prior original research articles (columns 2-6) or if they have no prior original research articles (column 1). All models incorporate a full suite of subfield and year fixed effects as well as a series of 9 indicator variables for subfield age. Exponentiated coefficients are presented, which after subtracting one yield magnitudes interpretable as elasticities. Robust errors in parentheses, clustered at the subfield ([†] $p < 0.10$, * $p < 0.05$, ** $p < 0.01$).

Table A4b. Effect of CPG by within subfield authors

	Authors new to subfield	Authors previously active in subfield
Treated $\times$ AfterCPG	1.069** (0.016)	1.193** (0.035)
Log pseudolikelihood	-709,650	-336,318
Number of root papers	16,524	15,621
Observations	333,403	315,015

Note: Estimates stem from conditional quasi-maximum likelihood Poisson regressions. The dependent variables are the number of original research articles in the subfield by whether the last author has previously published within the subfield. All models incorporate a full suite of subfield and year fixed effects as well as a series of 9 indicator variables for subfield age. Exponentiated coefficients are presented, which after subtracting one yield magnitudes interpretable as elasticities. Robust errors in parentheses, clustered at the subfield ([†] $p < 0.10$, * $p < 0.05$, ** $p < 0.01$).

Table A5. Dose effect

	Root article level	Subfield level			
	Citations	Article count	Citations	Ever cited by patent	Translational research
Cited by 1 CPG	1.050* (0.022)	1.069** (0.021)	1.091** (0.018)	1.101** (0.033)	1.274** (0.060)
Cited by 2 CPGs	1.234** (0.045)	1.116** (0.034)	1.181** (0.033)	1.239** (0.055)	1.342** (0.104)
Cited by 3 CPGs	1.272** (0.057)	1.167** (0.045)	1.187** (0.046)	1.466** (0.102)	1.496* (0.303)
Cited by 4 or more CPGs	1.958** (0.143)	1.352** (0.057)	1.562** (0.070)	1.724** (0.118)	1.412* (0.217)
Log pseudolikelihood	-340,570	-808,552	-4,332,529	-145,966	-83,720
Number of root papers	16,169	16,526	16,549	13,667	8,250
Observations	228,642	333,442	305,265	275,988	165,423

Note: Each column contains an estimate from a regression of the dependent variable in the column header upon a set of 4 indicator variables interacting post-treatment with the number of different CPGs citing a root article. All estimates stem from conditional quasi-maximum likelihood Poisson regressions. Column one is performed at the level of the root article, while in the remaining columns the analysis is performed at the level of the subfield. All models incorporate a full suite of article or subfield fixed effects, year fixed effects, and a series of 9 indicator variables for article or subfield age. Exponentiated coefficients are presented, which after subtracting one yield magnitudes interpretable as elasticities. Robust errors in parentheses, clustered at the root article (column 1) or subfield (columns 2-5) ($^{\dagger}p < 0.10$, $^*p < 0.05$, $^{**}p < 0.01$).

Table A6. Placebo tests

	Placebo year: Pre-treatment		Placebo year: Dropped from guideline	
	Root article citations	Subfield article count	Root article citations	Subfield article count
Treated $\times$ AfterCPG	1.033 (0.022)	1.029 (0.020)	0.984 (0.042)	0.975 (0.028)
Log pseudolikelihood	-132,834	-331,330	-132,039	-238,963
Number of root papers	10,737	11,112	5,839	6,050
Observations	92,200	154,004	92952	119728

Note: Estimates stem from conditional quasi-maximum likelihood Poisson regressions. The dependent variable is citations to the root article (columns 1 and 3) and flow of articles into the subfield (columns 2 and 4). The first two columns present the results with a randomly assigned treatment year during the period before inclusion in the CPG. The last two columns estimate the treatment effect for articles that were dropped from a CPG when it was updated; the treatment year is the year they were dropped. All models incorporate a full suite of article or subfield fixed effects, year fixed effects, and a series of 9 indicator variables for article or subfield age. Exponentiated coefficients are presented, which after subtracting one yield magnitudes interpretable as elasticities. Robust errors in parentheses, clustered at the root article (columns 1 and 3) or subfield (columns 2 and 4) level ($^{\dagger}p < 0.10$, $^*p < 0.05$, $^{**}p < 0.01$).

Table A7a. Effect of CPG inclusion on root article citations

	Citations	Author citations		Journal citations	
		Previously cited	New	Previously cited	New
Treated $\times$ AfterCPG	1.145** (0.023)	1.187** (0.037)	1.158** (0.022)	1.143** (0.028)	1.120** (0.017)
Log pseudolikelihood	-341,176	-445,891	-1,294,950	-169,257	-257,714
Number of root papers	16,169	13,342	16,164	13,352	16,132
Observations	228,642	188,417	228,562	190,241	227,992

Note: Estimates stem from conditional quasi-maximum likelihood Poisson regressions. The dependent variables are listed in the columns. All models incorporate a full suite of root article and year fixed effects as well as a series of 9 indicator variables for root article age. Exponentiated coefficients are presented for Poisson specifications, which after subtracting one yield magnitudes interpretable as elasticities. Robust errors in parentheses, clustered at the root article ($^{\dagger}p < 0.10$, $^*p < 0.05$, $^{**}p < 0.01$).

Table A7b. Effect of CPG inclusion on author and journal diversity

	Diversification index	
	Authors	Journals
Treated $\times$ AfterCPG	0.008** (0.002)	0.018** (0.003)
R ² (within)	0.0153	0.0446
R ² (overall)	0.0573	0.0501
Number of root papers	16,167	16,169
Observations	140,484	140,564

Note: Estimates are fixed effect linear regression. The dependent variable is the diversification index for authors (column 1) and journals (column 2) proposed by Teodoridis et al. (2019). All models incorporate a full suite of root article and year fixed effects as well as a series of 9 indicator variables for root article age. Robust errors in parentheses, clustered at the root article ($^{\dagger}p < 0.10$, $^*p < 0.05$, $^{**}p < 0.01$).

Table A8. Alternative control groups

	Root article level		Subfield level		
	Citations	Article count	Citations	Ever cited by patent	Translational research
Base case	1.145** (0.023)	1.093** (0.017)	1.134** (0.017)	1.197** (0.031)	1.213** (0.049)
Base case excluding any with shared MeSH	1.178** (0.029)	1.086** (0.026)	1.143** (0.022)	1.178** (0.042)	1.343** (0.072)
Base case limited to exact MeSH matches	1.147** (0.046)	1.119** (0.036)	1.129** (0.033)	1.255** (0.062)	1.252** (0.106)
CEM: Pulmonary disease	1.154** (0.021)	1.027 [†] (0.014)	1.078** (0.014)	1.032 (0.025)	1.150** (0.040)
CEM: Exact CV MeSH	1.218** (0.026)	1.107** (0.020)	1.163** (0.020)	1.209** (0.036)	1.198** (0.058)
Nearest neighbors	1.229** (0.019)	1.136** (0.012)	1.176** (0.011)	1.152** (0.020)	1.239** (0.036)

Note: Each cell contains an estimate for the treatment effect in a separate regression, with the analysis performed at the level of the root article (column 1) or subfield (columns 2-5). Estimates all stem from conditional quasi-maximum likelihood Poisson regressions. Each row estimates the treatment effect using a different treated and control group. The dependent variables are listed in the columns. All models incorporate a full suite of article or subfield fixed effects, year fixed effects, and a series of 9 indicator variables for article or subfield age. Exponentiated coefficients are presented, which after subtracting one yield magnitudes interpretable as elasticities. Robust errors in parentheses, clustered at the root article (columns 1) or subfield (columns 2-5) level ([†] $p < 0.10$, * $p < 0.05$, ** $p < 0.01$). CEM—coarsened exact matching.

Table A9. Exclusion of articles citing the CPG or root article

	Root article level		Subfield level	
	Citations	Excluding those citing treating CPG	Article count	Excluding those citing root article
Treated $\times$ AfterCPG	1.145** (0.023)	1.104** (0.024)	1.095** (0.019)	1.073** (0.017)
Log pseudolikelihood	-341,176	-336,493	-808,788	-691,312
Number of root papers	16,169	16,094	16,526	16,523
Observations	228,642	227,712	333,442	304,842

Note: Estimates stem from conditional quasi-maximum likelihood Poisson regressions. The dependent variable is citations to root article (columns 1 and 2) and flow of articles into the subfield (columns 3 and 4). Column 2 excludes from the dependent variable those citations that also cite the CPG that treated the root article, while column 4 excludes those subfield articles that cite the root article. All models incorporate a full suite of article or subfield fixed effects, year fixed effects, and a series of 9 indicator variables for article or subfield age. Exponentiated coefficients are presented, which after subtracting one yield magnitudes interpretable as elasticities. Robust errors in parentheses, clustered at the root article (columns 1 and 2) or subfield (columns 3 and 4) level ($^{\dagger}p < 0.10$, $^*p < 0.05$, $^{**}p < 0.01$).

Table A10. Alternative functional forms

	Root article level		Subfield level		
	Citations	Article count	Citations	Ever cited by patent	Translational research
Treated $\times$ AfterCPG	0.104** (0.010)	0.069** (0.012)	0.101** (0.015)	0.054** (0.005)	0.033** (0.003)
R ² (within)	0.0462	0.0844	0.4147	0.0914	0.0102
R ² (overall)	0.0088	0.0242	0.0917	0.0661	0.0052
Number of root papers	16,550	16,550	16,550	16,550	16,550
Observations	234,534	333,906	305,286	333,906	333,906

Note: Estimates all stem from fixed effect linear regression. The dependent variables are inverse hyperbolic sine of the column headings. All models incorporate a full suite of article or subfield fixed effects, year fixed effects, and a series of 9 indicator variables for article or subfield age. Robust errors in parentheses, clustered at the root article (columns 1) or subfield (columns 2-5) level ($^{\dagger}p < 0.10$, $^*p < 0.05$, $^{**}p < 0.01$).

Table A11. Two-period OLS

	Root article level	Subfield level
	Citations	Article count
Treated $\times$ AfterCPG	0.732** (0.0928)	0.241** (0.085)
R ²	0.010	0.009
Number of root papers	16,550	16,550
Observations	32,806	33,100

Note: The original panel data is collapsed into two time periods, pre- and post-treatment. The estimate stems from OLS regression of mean annual citations to the root article (column 1) or mean annual count of original research articles within the subfield (column 2) on an indicator variable for post-treatment, an indicator variable for if received CPG inclusion, and the interaction of the two. Robust errors in parentheses, clustered at the root article (columns 1) or subfield (column 2) level ($^{\dagger}p < 0.10$, $^*p < 0.05$, $^{**}p < 0.01$).

Table A12. Analysis at level of CPG

	Root article level	Subfield level
	Citations	Article count
Treated $\times$ AfterCPG	1.284* (0.126)	1.081 (0.054)
Log pseudolikelihood	-78,895	-76,158
Number of guidelines	254	254
Observations	4,655	5,141

Note: Estimates stem from conditional quasi-maximum likelihood Poisson regressions. The dependent variable is citations to the root article (columns 1) and count of articles within the subfield (columns 2). All models incorporate a full suite of article or subfield fixed effects, year fixed effects, and a series of 9 indicator variables for article or subfield age. The number of guidelines reflects both CPGs citing treated articles and “counterfactual CPGs” of control articles, inherited from the matched treated article. Exponentiated coefficients are presented, which after subtracting one yield magnitudes interpretable as elasticities. Robust errors in parentheses, clustered at the root article (columns 1) or subfield (column 2) level ($^{\dagger}p < 0.10$, $^*p < 0.05$, $^{**}p < 0.01$).

Table A13. Stacked regression

	Root article level	Subfield level			
	Citations	Article count	Citations	Ever cited by patent	Translational research
Base case	0.104** (0.010)	0.069** (0.012)	0.101** (0.015)	0.054** (0.005)	0.033** (0.003)
Stacked regression	0.109** (0.010)	0.055** (0.011)	0.089** (0.013)	0.049** (0.005)	0.034** (0.004)

Note: Each cell contains an estimate for the treatment effect in a separate regression, with the analysis performed at the level of the root article (column 1) or subfield (columns 2-5). All estimates stem from fixed effect linear regression. The dependent variables are inverse hyperbolic sine of the column headings. Row 1 repeats the base case analysis from Table A10. Row 2 implements stacked regression following Cengiz et al. 2019; see text above for details. Robust errors in parentheses, clustered at the root article (columns 1) or subfield (columns 2-5) level ([†] $p < 0.10$, * $p < 0.05$, ** $p < 0.01$).

Appendix B: Clinical Practice Guidelines

CPGs have a wide range in their scope and topics covered. For instance, they may include the relatively narrow scope of the British Society for Rheumatology’s CPG on the use of TNF-alpha blockers in ankylosing spondylitis to the broad scope of the American Psychiatry Association’s CPG on depression. Similarly, they range from the Infectious Disease Society of America’s CPG for the relatively benign strep throat to the life threatening American College of Cardiology/American Heart Association’s CPG on heart attacks.

To fix ideas, consider the 2005 American College of Cardiology/American Heart Association (ACC/AHA) CPG on heart failure (Hunt et al. 2005). It has 4 specific recommendations regarding which heart failure patients should receive angiotensin converting enzyme inhibitors, such as “Angiotensin converting enzyme inhibitors are recommended for all patients with current or prior symptoms of heart failure and reduced left ventricular ejection fraction, unless contraindicated” with a level of evidence of “A,” meaning it is supported by multiple high quality randomized controlled trials. The discussion of angiotensin converting enzyme inhibitors cites over 50 supporting references. It includes discussion of the underlying basic science (“It is not clear whether the effects. . . can be explained solely by the suppression of angiotensin II production, because [angiotensin converting enzyme] inhibition. . . also enhances the action of kinins and augments kinin-mediated prostaglandin production. In experimental models of heart failure. . .”) as well as specific, practical recommendations for their use (“Renal function and serum potassium should be assessed within 1 to 2 weeks of initiation of therapy”). In contrast, the same CPG states “Physicians should use diuretics to control pulmonary congestion and peripheral edema in patients with heart failure and normal left ventricular ejection fraction” with a level of evidence of “C,” meaning it is supported by expert opinion or methodologically limited observational studies.

While there is variation across professional societies in the processes used to construct CPGs, over time, these processes have become increasingly larger in scope, complexity, and formalization. Jacobs et al. (2014) and Levine et al. (2019) provide a detailed description of the process used by the ACC/AHA to develop CPGs. To contextualize the discussion of CPGs, I will briefly review the ACC/AHA drafting process as of 2014 as an illustrative example. The ACC/AHA Task Force on Clinical Practice Guidelines selects the membership of a Guideline Writing Committee (GWC) for that particular disease. The GWC performs detailed evidence acquisition, reviews it, and drafts recommendations based on the strength and quality of evidence with a limited preliminary text. Their work is supplemented by a separate Evidence Review Committee (ERC). This committee, whose membership includes methodologists, epidemiologists, and biostatisticians in addition to clinicians, is tasked with performing formal systematic reviews, including assessing the quality of the underlying evidence. The ERC focuses on those questions where the evidence and clinical relevance suggests it would be most informative. The recommendations and text are then updated by the GWC using the evidence reviewed by the ERC. In areas where recommendations may overlap and potentially be discordant with other ACC/AHA CPGs for related disease processes, the CPG enters a reconciliation process where it receives input from these other groups and may result in sections jointly drafted across multiple CPGs. Each recommendation is ratified by a formal vote of the GWC membership. The updated document then undergoes official peer review by ACC and AHA reviewers, content experts, and outside collaborators and stakeholders where appropriate. Received comments, which may number over one thousand, are reviewed and incorporated by the GWC, with any changes in recommendations requiring a formal vote. The final document is then ratified by the ACC Board of Trustees and AHA Science Advisory and Coordinating Committee.

While other organizations sponsored CPGs, including insurance companies, individual hospitals, and non-profits, the CPGs from professional medical societies became particular prominent. In particular, these societies typically had a national role and connected with medicine’s tradition of autonomy and self-regulation. Professional society guidelines, given the professional norm of placing patient interests first and the rigorous process under which they are typically developed, were frequently seen as more authoritative than CPGs produced by insurance companies or hospitals (Mello 2001).

References

- Jacobs, Alice K., Jeffrey L. Anderson, Jonathan L. Halperin. 2014. "The evolution and future of ACC/AHA clinical practice guidelines: a 30-year journey: A report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines." *Journal of the American College of Cardiology*, 64(13):1373-84.
- Levine, Glenn N., Patrick T. O'Gara, Joshua A. Beckman, Sana M. Al-Khatib, Kim K. Birtcher, Joaquin E. Cigarroa, Lisa de Las Fuentes, Anita Deswal, Lee A. Fleisher, Federico Gentile, Zachary D. Goldberger, Mark A. Hlatky, José A. Joglar, Mariann R. Piano, Duminda N. Wijeyesundera. 2019. "Recent Innovations, Modifications, and Evolution of ACC/AHA Clinical Practice Guidelines: An Update for Our Constituencies: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines." *Journal of the American College of Cardiology*, 73(15):1990-1998.

Appendix C: Measurement and Methods

Coarsened exact matching

I employ coarsened exact matching (CEM), a non-parametric method of selecting control observations. Selecting covariates on which to match involves some judgment, but is driven by my belief in which variables may be related to both receiving treatment and outcomes of interest. Rather than attempt to match against a continuous variable, covariates are first coarsened into discrete ranges. A set of strata is then created, corresponding to the entire support of the joint distribution of the coarsened covariates. Each treated and potential control observation is then assigned to stratas, with controls selected from the same strata as the treated observation. I implement one-to-one matching given a single article may serve as a control at multiple points in time during the study period. In implementing CEM, there is a trade-off: the better the balance achieved by adding additional covariates to match against or using narrower coarsening windows, the less statistical power is available for inference as a greater number of treated observations will fail to share a strata with any potential control article.

I implement CEM by matching the 14,930 root articles against 831,218 potential controls, representing all original research articles in PubMed, in English, published 2011 or earlier, with at least 1 CV disease MeSH term. For three sets of variables, I impose an exact match: publication year; a set of 4 indicator variables for research methodology (i.e., clinical trial, case report, epidemiologic, and bench science); and a set of 16 indicator variables for CV disease topic. I also match on five coarsened variables. I first match on pre-treatment cumulative citations from original research articles using 7 coarsened bins: 0, 1 to 10, 11 to 30, 31 to 60, 61 to 100, 101 to 200, and 201 and above. I also match on citations in the year prior to treatment: 0, 1 to 5, 6 to 20, and 21 and above. While each treated root article has a fixed publication year and treatment year, controls have a fixed publication year but inherit a counterfactual “mock treatment year” from the treated root article to which they are matched. Since a control article not selected in a given year may serve as a control in the next year, pre-treatment citations are tabulated separately for each year of potential treatment. Journal impact factor is coarsened in the following five bins: 0 to 2, >2 to 5, >5 to 10, >10 to 20, and >20. As a measure of distance to the scientific frontier, I look at the mean age, in years, of original research references. This is coarsened into five bins: 0 to 4, >4 to 11, >11 to 20, and >20. Finally, as a measure of the breadth of knowledge a paper draws upon, I match on the coarsened number of original research references: 0, 1 to 9, 10 to 30, 31 to 50, and 51 and above.

This process results in a match for 8,275 (55.4%) of treated root articles to 8,275 controls articles. Since the selection of covariates and thresholds for coarsening involves some subjectivity, as a robustness check, this process was repeated with variations in the number of coarsening bins; thresholds used in coarsening; and with additional covariates including number of authors, exact matches to journal, and additional methodology controls with similar results obtained.

Alternative control groups

As a robustness analysis, I repeat my analysis using three alternative control groups. Descriptives and covariate balance are displayed in Appendix Table A2 and the full distribution of CV disease categories is available in Appendix Table C2. First, I construct a control group very tightly matched on disease topic with treated articles. To do this, I identify the frequency of each CV disease MeSH term among the entire population of potential controls. I then repeat the CEM process exactly as outlined above, but instead of matching on 16 coarsened CV disease categories, I require an exact match between the least common CV MeSH term of each paper. In this way, I ensure overlap in the most specific, narrow disease topic addressed by the papers. This results in a match for 6,360 (42.6%) of treated articles.

To construct a control group which is more distantly related with treated articles, such that there is not a concern a control article might be adversely impacted by its exclusion from a CV CPG, I use articles tagged with a pulmonary disease MeSH as a control. To ensure I am matching diseases of similar relative importance, I calculate the percentile of each MeSH term’s frequency among all original research articles in PubMed, 2011 and before, tagged with at least one CV disease MeSH or one pulmonary disease MeSH, respectively. I then repeat the CEM process exactly as outlined above, but instead of matching on 16 coarsened CV disease categories, I instead match on the coarsened MeSH frequency percentile: 0 to 25, >25 to 50, >50 to 75, >75 to 90, and >90. This results in a match for 10,262 (68.7%) of treated articles.

Finally, to allow me to estimate a treatment effect for the full sample of treated articles, I use nearest neighbor articles as controls. I first identify the set of all articles in the same specific journal issue as the 14,930 root articles. I keep from this set only those which were original research and in English. I also exclude articles which, at the time of counterfactual treatment of the root article, were themselves previously included in a CPG of any medical specialty by medical professional societies in the United States, United Kingdom, Canada, and Australia and New Zealand. For each root article, I identify the two papers closest in distance as measured by the number of pages. If the second and third closest articles are of equal distance, an article is randomly selected as the control. In some cases, only a single eligible neighboring article is present; in these cases only a single control is included for that particular root article. This results in 27,728 nearest neighbor control articles.

Commercial orientation: author industry affiliation

I used author affiliation in Web of Science to measure whether the article has any industry-affiliated authors; this data is only available through 2014. Authors were considered to be from industry if the affiliation had one of the following terms: corporation, incorporated, llc, company, venture, limited, pharma dev, chem co, technology co. Care was taken to measure derivative terms (e.g. “corp” for corporation). The resulting list was carefully reviewed and any organizations not corresponding to private industry were removed (e.g., “mitre corp” and “Copenhagen hosp corp”). This was supplemented by directly searching for the top 100 medical device, pharmaceutical, and biotechnology companies. Examples affiliations added via this approach include “Celgene,” “Medtronic,” and “Merck.”

Research methodology and topic

To characterize research methodology and topic, I capitalize on Medical Subject Heading (MeSH), a hierarchical controlled vocabulary, created and maintained by the National Library of Medicine used for indexing and searching the biomedical literature. In the 2019 edition of MeSH, there are over 29,000 descriptors, with an average of about 12 MeSH terms associated with each publication. Crucially, MeSH terms are assigned by professional indexers rather than by authors who might have an incentive to strategically select MeSH terms (Bachrach and Charen 1978). While it is difficult to assess accuracy as there is not a single “correct” way to index articles, MeSH have generally been found to be consistent across indexers (Coletti and Bleich 2001, Funk and Reid 1983). For some measures, MeSH was supplemented with title keywords and PubMed publication type category. In all cases, these measures were used only to describe original research articles, and thus, do not represent review articles, meta-analysis, letters and other such publications.

Clinical trial

To measure clinical trial research, I combine three different measures, with an emphasis on capturing phase 1-4 trials. I start with using the following MeSH codes: V03.175.250.100, E05.318.372.250.250.200, V03.175.250.200, E05.318.372.250.250.210, V03.175.250.300, E05.318.372.250.250.220, V03.175.250.400, E05.318.372.250.250.230, and E05.337.800.200. This identifies 8 distinct MeSH terms, with examples including *phase 2 clinical trial* and *phase 4 clinical trial*. I supplement this approaching using the four article publication types corresponding to phase 1-4 clinical trials (e.g., *clinical trial, phase 1*). Finally, articles with titles including the phrases “phase 1,” “phase 2,” “phase 3,” or “phase 4” were also included. These were reviewed to exclude topics that might be picked up by these search terms but were unrelated to clinical trials, such as those related to bundle branch blocks, blood pressure, and cell depolarization.

Epidemiologic

Epidemiologic studies were characterized by the following MeSH codes as well as all codes under them in the hierarchical tree: E05.318.372.500, N05.715.360.330.500, and N06.850.520.450.500. In total, this identifies 12 MeSH terms. Examples include *cohort studies*, *case-control studies*, and *seroepidemiologic studies*. Care was taken to exclude terms reflecting a specific study or patient population rather than a general epidemiologic method, such as excluding the MeSH for *National Longitudinal Study of Adolescent Health* and *HIV Seroprevalence*. This measure was expanded to also include any papers with one the following title key words: “case-control,” “retrospective,” and “cohort.”

Case reports

Case reports were identified by publications tagged with the MeSH code V03.100 corresponding to the single MeSH term *case reports [publication type]*. Articles were also considered case reports if their title included the phrase “case report” or if they were the publication type *case reports*.

Bench science

To identify bench science articles, I followed the measure of Azoulay et al. (2021). This is comprised of four separate components. The first two focus on the tools of bench science, identifying articles using molecular biology methods or model organisms. The second two focus on broad topics distinct to molecular biology research: work on cellular structures and macromolecules and articles related to biochemical and cellular processes. An article is considered bench science if it has at least one MeSH term in any of these four categories.

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 that 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 capture use of model organisms, I compiled a list of 53 different model organisms used in biomedical research: *tobacco mosaic virus*, *bacteriophage lambda*, *bacteriophage phiX174*, *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 that 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.

I constructed two additional measures of bench science based on the research topic: the first focused on cellular structures and macromolecules, and the second on biochemical and cellular processes. I 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 that 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 captured 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*.

Cardiovascular and pulmonary disease

I considered any paper tagged under C14, *cardiovascular diseases*, in the MeSH hierarchy to be about cardiovascular disease. This results in 425 unique MeSH terms, with examples ranging from common diseases such as *myocardial infarction* or *atrial fibrillation* to the more esoteric *Lutembacher syndrome* or *Takotsubo cardiomyopathy*. To develop a coarser measure of cardiovascular disease topic, I constructed 16 cardiovascular disease categories. The 75 CV MeSH terms at the third level of the MeSH tree were then assigned, along with all MeSH under them in the hierarchical tree, to one of these 16 categories or to an other category. In the event a MeSH term was listed under multiple hierarchical third-level MeSH codes, the term was assigned to the most relevant overall category. Importantly, since papers may be tagged with multiple CV disease MeSH, an individual paper may belong to multiple of these 16 coarsened CV disease categories. The categories and corresponding MeSH codes are:

Aneurysms and aortic disease: C14.280.358, C14.907.055 (except C14.907.055.635, C14.907.055.050.150), C14.907.109 (except C14.907.109.239.650, C14.907.109.320)
Arrhythmias: C14.280.067, C14.280.123
Cardiac arrest: C14.280.383, C14.907.725 (except C14.907.725.600)
Cardiovascular infections: C14.260.249, C14.260.750 (except C14.260.750.595), C14.260.500, C14.280.282, C14.260
Congenital heart disease: C14.240.400 (except C14.240.400.145, C14.240.400.660), C14.280.400 (except C14.280.400.145, C14.280.400.660)
Heart failure and cardiomyopathies: C14.280.148, C14.280.195, C14.280.238 (except C14.280.238.615), C14.280.434, C14.280.832, C14.280.945
Hypertension; C14.907.489, C14.907.653, C14.907.690
Hypotension: C14.907.514, C14.907.946
Myocardial infarction: C14.280.470, C14.280.647 (except C14.280.647.250.250, C14.280.647.250.260), C14.280.671, C14.907.585 (except C14.907.585.250.250, C14.907.585.250.260)
Neurovascular disease: C14.907.253 (except C14.907.253.329.600, C14.907.253.535.800, C14.907.253.946), C14.907.601, C14.907.760, C14.907.790
Pericardial heart disease: C14.280.155, C14.280.695, C14.280.720, C14.280.763, C14.280.793, C14.260.750.595
Peripheral vascular disease: C14.907.137 (except C14.907.137.126.372, C14.907.137.230, C14.907.137.427, C14.907.137.615), C14.907.617 (except C14.907.617.249, C14.907.617.648, C14.907.617.718), C14.907.903, C14.907.927, C14.907.952 (except C14.907.952.880)
Thromboembolism: C14.583.404, C14.907.355 (except C14.907.355.590, C14.907.355.830), C14.907.940.740.910, C14.907.952.880
Valvular heart disease: C14.280.104, C14.280.484 (except C14.280.484.845, C14.280.484.716.525), C14.280.874, C14.280.955, C14.907.798
Vascular malformations: C14.907.077, C14.907.780, C14.907.823 (except C14.907.823.225), C14.907.933 (except C14.907.933.110.500, C14.907.933.555.500), C14.240.850 (except C14.240.850.750.147.500, C14.240.850.750.295, C14.240.850.937, C14.240.850.875.500, C14.240.850.984.500.500, C14.240.850.984.750.500)
Vasculitis: C14.907.940 (except C14.907.940.090.170, C14.907.940.530, C14.907.940.740.910, C14.907.940.890, C14.907.940.905, C14.907.940.907.175, C14.907.940.907.850)

Pulmonary diseases were similarly identified as all those under the MeSH code C08, *respiratory tract diseases*. This results in 238 unique MeSH terms, with examples including *interstitial lung diseases*, *pneumothorax*, and *mycoplasma pneumonia*.

Incremental clinical studies

I identify three categories of papers using MeSH to construct a measure of incremental clinical studies. I limit eligible papers to those including a disease MeSH (any MeSH C1-C21, C23-C26, and F3; this excludes C22 veterinary diseases) and no bench science MeSH.

The first subset of papers concerns populations under study. This comprises 22 MeSH capturing different ages and genders. Examples include *adolescent*, *pregnant woman*, *aged 80 and over*, and *men*. The final list of MeSH is: M01.060, M02.390, M01.975 (except M01.390.660, M01.975.155, M01.975.825, M01.975.790, M01.975.310).

I also measure different geographic locations in which a study can occur. This is captured by all MeSH under Z01 and corresponds to 401 distinct MeSH terms. Examples include *China*, *Kenya*, *Eastern Europe*, and *Massachusetts*.

Direct patient care

I also capitalize on MeSH to measure research tightly connected with direct patient care. I develop five separate measures, each capturing a distinct type of direct patient care.

The first measure captures physician clinical skills involved in direct patient care, including clinical reasoning, history taking, and physical exam skills. There are 22 distinct MeSH terms, with examples including *differential diagnosis*, *heart auscultation*, *clinical decision-making*, and *percussion*. Care was taken to exclude terms which might overlap with disease symptoms or management such as abnormal gaits and blood pressure determination. Terms associated with alternative care were also excluded. The final list of Mesh is: 01.390, E01.055, E01.171, E01.370.510, and E01.370.600. The following MeSH were excluded: E01.370.510.150, E01.370.510.700, E01.370.600.024, E01.370.600.050, E01.370.370.140, E01.370.600.100, E01.370.600.115, E01.370.600.225, E01.370.600.230, E01.370.600.250, E01.370.370.380.425, E01.370.600.315, E01.370.600.337, E01.370.600.425, E01.370.600.620, E01.370.600.700, E01.370.600.750, E01.370.600.875.

The second approach measures diagnostic tests. There are 708 distinct MeSH terms, with examples including *coronary angiography*, *blood culture*, *liver function tests*, and *frozen sections*. The final list of MeSH is: E01.390, E01.158, E01.370.685, E01.110, E02.760.273, E01.370 (excluding E01.370.620, E01.370.760, E01.370.505, E01.370.872, E01.370.503, E01.370.928, E01.370.510).

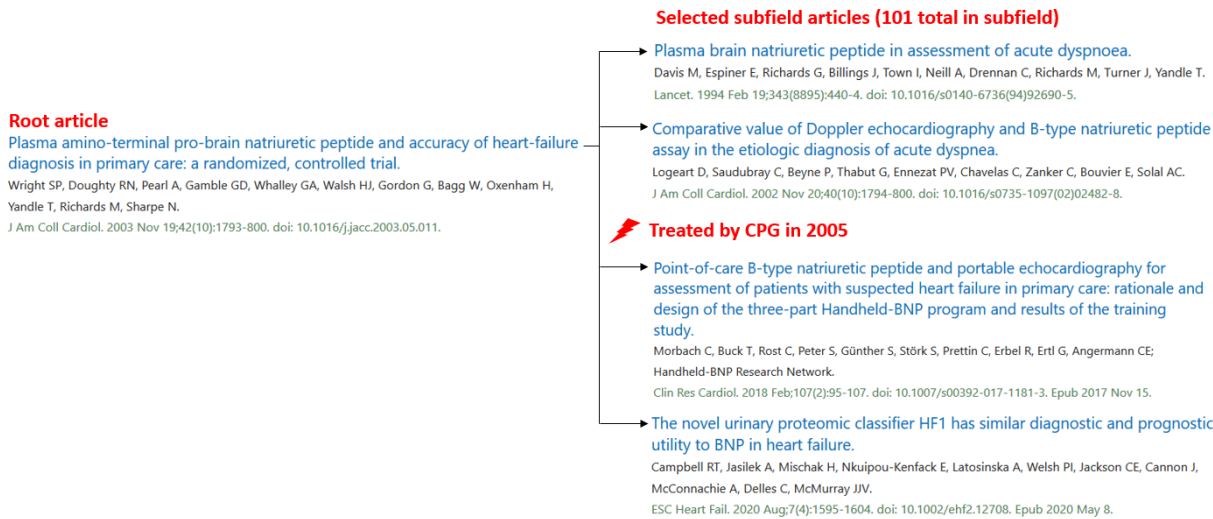
As a third measure, I identify those MeSH related to operative procedures. There are 584 distinct MeSH terms in this category, with examples including *thoracotomy*, *laparoscopy*, and *coronary balloon angioplasty*. This final list of MeSH is: E04 and E03 (excluding E03.091).

The fourth measures identifies clinical therapeutics. It includes 270 distinct MeSH terms, with examples such as *positive pressure respiration*, *fluid therapy*, *plasmapheresis*, and *premedication*. The final list of MeSH is: E02.041, E02.075, E02.148, E05.157, E02.154, E04.014.170, E02.950, E02.960, E02.926, E02.912, E02.897, E02.895, E02.891, E02.887, E02.880, E02.870, E02.860, E02.815, E02.808, E02.800, E02.278, E02.309, E02.467, E02.596, E02.600, E02.621, E02.631, E02.258, E02.514, E02.520, E02.585, E02.587, E02.592, E02.774, E02.365.647, E02.331, E02.779.468, E02.831.535.468, E02.095.125, E02.095.135, E02.120, E02.186, E02.197. This measure excludes MeSH associated with pharmacological actions (Category R), since these may have substantial overlap with basic pharmacologic research. The base case for this measure also includes MeSH with a connection to pharmacologic therapy (This includes E02.095 and E02.319, but excludes the following MeSH: E02.319.890, E02.900.890, E02.319.900, E02.900.900, E02.319.529, E02.183.750, E02.319.077, E02.319.490, E02.319.698, E02.319.306, E02.319.305, E02.319.300, E02.319.283, E02.319.267, E02.319.243). As a robustness check, the results were repeated excluding the 80 corresponding MeSH terms.

The final measures seeks to capture direct health care provision, typically with terms related to quality improvement. There are 188 distinct MeSH terms, with examples including *case management*, *patient compliance*, *health care quality indicators*, and *patient outcome assessment*. Care was taken to avoid any MeSH related to guidelines or standards themselves. The final list of MeSH is: N04.761, N04.590, N04.761.700, N05.700, N04.761.616, N05.715.680, N04.761.744, N04.590.374, N05.300, N04.761.210, N05.715.175, N05.715.360.700, N04.452.822.700, N05.300.150.800.625, N05.715.360.600, N04.590.233.727.210.249, N04.761.559, N05.715.360.575 (and excluding N04.761.337, N05.715.360.395, N04.761.789.900, N05.715.840, N04.761.700.350, N05.700.350).

Since some terms may appear in multiple of the above categories—for instance, some procedures can be both diagnostic and therapeutic—in total, these measures contain 1,618 distinct MeSH.

Figure C1. Subfield construction and evolution



Note: I illustrate the process of constructing subfields and tracing research trajectory through an example. The left-hand side panel presents the root article, a trial of a biomarker in diagnosing heart failure in the primary care setting. The PMRA is used to construct the subfield around the root article based on share textual similarity, selected examples of which can be seen in the right-hand side panel. I can then characterize the type of research within the subfield before and after the CPG.

Figure C2. Illustration of treated and control root articles

	Results of percutaneous transluminal valvuloplasty in 218 adults with valvular aortic stenosis	Long-term results of balloon pulmonary valvuloplasty of valvar pulmonic stenosis
	B Letac [†] , A Cribier, R Koning, J P Bellefleur	P S Rao [†] , M E Fawzy, L Solymar, M K Mardini
Treatment status	ACC/AHA CPG on valvular heart disease, 1998	Control
Publication year	1988	1988
Cumulative citations	40	35
Citations in prior year	1	1
Journal	American Journal of Cardiology	American Heart Journal
Journal impact factor	3.1	3.6
References	16	18
Subfield size	63	64
Subfield age (years)	20	23
Research methodology	Observational case series	Observational case series
Disease topic	Aortic stenosis + balloon valvuloplasty	Pulmonic stenosis + balloon valvuloplasty

Note: The two articles above illustrate the Coarsened Exact Matching (CEM) procedure. Each occur in the same year and are very similar along citations, journal impact factor, references, subfield size, subfield age, and research methodology. In line with coarsely matching on disease, both focus on valvular heart disease but with different valves of the heart.

Table C1. List of sponsoring organizations

Sponsoring Organization	Number of guidelines (%)
American Heart Association	75 (55.6%)
American College of Cardiology	73 (54.1%)
Canadian Cardiovascular Society	30 (22.2%)
Society for Cardiovascular Angiography and Interventions	18 (13.3%)
Heart Rhythm Society	15 (11.1%)
Hypertension Canada	11 (8.1%)
Cardiac Society of Australia and New Zealand	9 (6.7%)
Society for Vascular Medicine	9 (6.7%)
Society of Cardiovascular Anesthesiologists	9 (6.7%)
Society of Thoracic Surgeons	9 (6.7%)
American Society of Echocardiography	8 (5.9%)
American Society of Nuclear Cardiology	8 (5.9%)
Society for Vascular Surgery	8 (5.9%)
Society of Interventional Radiology	6 (4.4%)
American Association for Thoracic Surgery	4 (3.0%)
British Hypertension Society	4 (3.0%)
European Society of Cardiology	4 (3.0%)
American College of Radiology	3 (2.2%)
Heart Failure Society of America	3 (2.2%)
International Society for Heart and Lung Transplantation	3 (2.2%)
American College of Chest Physicians	2 (1.5%)
American College of Emergency Physicians	2 (1.5%)
American College of Physicians	2 (1.5%)
American Stroke Association	2 (1.5%)
Canadian Heart Rhythm Society	2 (1.5%)
Infectious Disease Society of America	2 (1.5%)
American Academy of Neurology	1 (0.7%)
American Academy of Pediatrics	1 (0.7%)
American Association of Neurological Surgeons	1 (0.7%)
American Association of Neuroscience Nurses	1 (0.7%)
American Dental Association	1 (0.7%)
American Geriatrics Society	1 (0.7%)
American Society for Gastrointestinal Endoscopy	1 (0.7%)
American Society for Preventive Cardiology	1 (0.7%)
American Society of Hypertension	1 (0.7%)
American Society of Nephrology	1 (0.7%)
American Society of Neuroradiology	1 (0.7%)
Association of Black Cardiologists	1 (0.7%)
British Cardiovascular Intervention Society	1 (0.7%)
Canadian Academy of Child and Adolescent Psychiatry	1 (0.7%)
Canadian Association of Nuclear Medicine	1 (0.7%)
Canadian Association of Radiologists	1 (0.7%)
Canadian Cardiac Magnetic Resonance	1 (0.7%)
Canadian Nuclear Cardiology Society	1 (0.7%)
Canadian Paediatric Society	1 (0.7%)
Canadian Society for Vascular Surgery	1 (0.7%)
Canadian Thoracic Society	1 (0.7%)
Congress of Neurological Surgeons	1 (0.7%)
European Heart Rhythm Association	1 (0.7%)
European Society of Hypertension	1 (0.7%)
International Society for Adult Congenital Heart Disease	1 (0.7%)
National Heart Foundation of Australia	1 (0.7%)
Royal College of Physicians	1 (0.7%)
Society for Cardiac Angiography and Interventions	1 (0.7%)
Society for Vascular Nursing	1 (0.7%)
Society of Atherosclerosis Imaging and Prevention	1 (0.7%)
Society of NeuroInterventional Surgery	1 (0.7%)

Note: Sponsoring organizations for the 135 guidelines under study. Organizations above that are not professional societies or are not cardiology organizations co-sponsored the guideline with a cardiology professional society.

Table C2. Distribution of cardiovascular disease categories across samples

Coarsened CV disease category	Potential CV controls	All treated articles	CEM: Base case	CEM: Pulm. disease	CEM: Exact CV MeSH	Nearest neighbor controls
Aneurysms and aortic disease	44,480 (5.4%)	1,082 (7.2%)	413 (5.0%)	820 (8.0%)	333 (5.2%)	1,078 (3.9%)
Arrhythmias	77,050 (9.3%)	2,841 (19.0%)	1,227 (14.8%)	1,960 (19.1%)	962 (15.1%)	2,943 (10.6%)
Cardiac arrest	27,079 (3.3%)	521 (3.5%)	92 (1.1%)	325 (3.2%)	161 (2.5%)	478 (1.7%)
Cardiovascular infections	10,623 (1.3%)	379 (2.5%)	69 (0.8%)	274 (2.7%)	96 (1.5%)	166 (0.6%)
Congenital heart disease	50,492 (6.1%)	971 (6.5%)	374 (4.5%)	656 (6.4%)	259 (4.1%)	1,343 (4.9%)
Heart failure and cardiomyopathies	88,468 (10.6%)	2,664 (17.8%)	1,211 (14.6%)	1,750 (17.1%)	1,085 (17.1%)	3,228 (11.7%)
Hypertension	96,342 (11.6%)	1,123 (7.5%)	513 (6.2%)	849 (8.3%)	618 (9.7%)	1,391 (5.0%)
Hypotension	12,514 (1.5%)	106 (0.7%)	8 (0.1%)	80 (0.8%)	36 (0.6%)	126 (0.5%)
Myocardial infarction	148,080 (17.8%)	5,625 (37.7%)	3,375 (40.8%)	3,738 (36.4%)	2,816 (44.3%)	6,496 (23.5%)
Neurovascular disease	130,746 (15.7%)	1,306 (8.7%)	500 (6.0%)	938 (9.1%)	479 (7.5%)	1,215 (4.4%)
Pericardial heart disease	8,691 (1.0%)	76 (0.5%)	13 (0.2%)	51 (0.5%)	9 (0.1%)	154 (0.6%)
Peripheral vascular disease	84,714 (10.2%)	1,520 (10.2%)	683 (8.3%)	1,120 (10.9%)	605 (9.5%)	1,808 (6.5%)
Thromboembolism	64,547 (7.8%)	624 (4.2%)	142 (1.7%)	426 (4.2%)	180 (2.8%)	698 (2.5%)
Valvular heart disease	38,432 (4.6%)	1,555 (10.4%)	517 (6.2%)	1,074 (10.5%)	442 (6.9%)	1,525 (5.5%)
Vascular malformations	18,503 (2.2%)	17 (0.1%)	1 ($<0.1\%$)	9 (0.1%)	0 (0.0%)	94 (0.3%)
Vasculitis	22,481 (2.7%)	61 (0.4%)	10 (0.1%)	40 (0.4%)	12 (0.2%)	147 (0.5%)
Other CV disease	87,189 (10.5%)	422 (2.8%)	298 (3.6%)	297 (2.9%)	207 (3.3%)	660 (2.4%)
Total number of papers	831,218	14,930	8,275	10,262	6,360	27,651

Note: Column percentages in parenthesis. Column 1 is comprised of all original publications, indexed in PubMed, with at least one cardiovascular disease MeSH. Column 2 is all original publications with a CV disease MeSH referenced by the 135 guidelines under study. Column 3-5 is identical to column 2, but includes only those publications which were matched to controls via coarsened exact matching using coarsened CV disease categories (column 3), CV MeSH frequency compared to pulmonary disease articles (column 4), and matching exact CV MeSH terms (column 5). Column 6 is the population of nearest neighbors to column 2.

Chapter 3

Long-term Effects from Early Exposure to Research: Evidence from the NIH “Yellow Berets”*

(with Pierre Azoulay and Misty Heggeness)

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.

* This chapter has been published by Elsevier and is reproduced here, with permission, with minor formatting changes. The full citation for the article is: Azoulay, Pierre, Wesley H. Greenblatt, and Misty L. Heggeness. 2021. “Long-term Effects from Early Exposure to Research: Evidence from the NIH “Yellow Berets.” *Research Policy*, 50(9):104332. This work was supported by the National Institutes of Aging (grant number R24-AG048059) to the National Bureau of Economic Research.

“[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 69th 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

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.

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 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

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.

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

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.

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 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-

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.

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 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

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.

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).

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

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.

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 ($\Lambda\Omega\Lambda$),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 gone out into*

⁷ Criteria for selection into $\Lambda\Omega\Lambda$ 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.

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 applicant’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 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

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

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 Tenreiro 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 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 ΑΩΑ 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

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 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 ($p < 0.01$).

abundance of explanatory variables in these models. Our least restrictive specification includes 94 fixed effects for medical schools and 238 indicator 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 naive 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 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.

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.

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;

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.

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-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 50th, above the 75th, above the 95th, above the 99th, and above the 99.9th 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.

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. 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.

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, 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

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.

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.

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 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,

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.

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 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 21st 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 5th and 95th percentiles of the distribution of lasso weights; and winsorization at the 10th and 90th 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 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.

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.

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 vitae (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 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

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; 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).

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.

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,

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).

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.

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 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

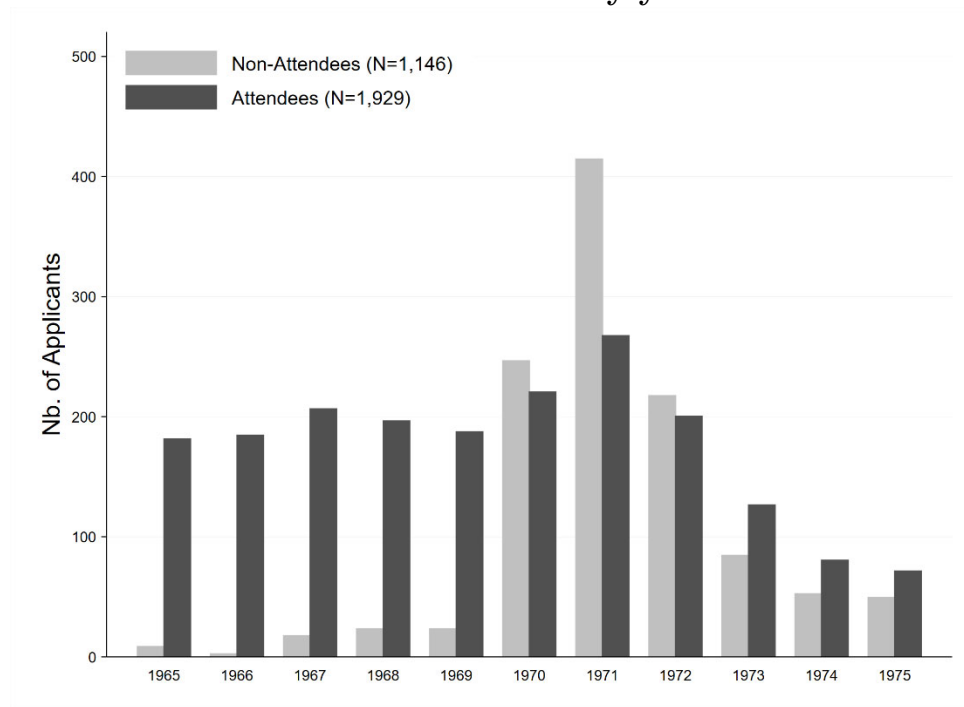
- Aghion, Philippe, and Peter Howitt. 1992. "A Model of Growth through Creative Destruction." *Econometrica*, 60(2): 323-351.
- Association of American Medical Colleges. 2016. "Diversity in Medical Education: Facts & Figures 2016." Washington, D.C.: AAMC.
- Anfinsen, Christian B. 1963. "History of the 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 Nathaniel 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: Nondiagnostic 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 Tenreiro. 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.

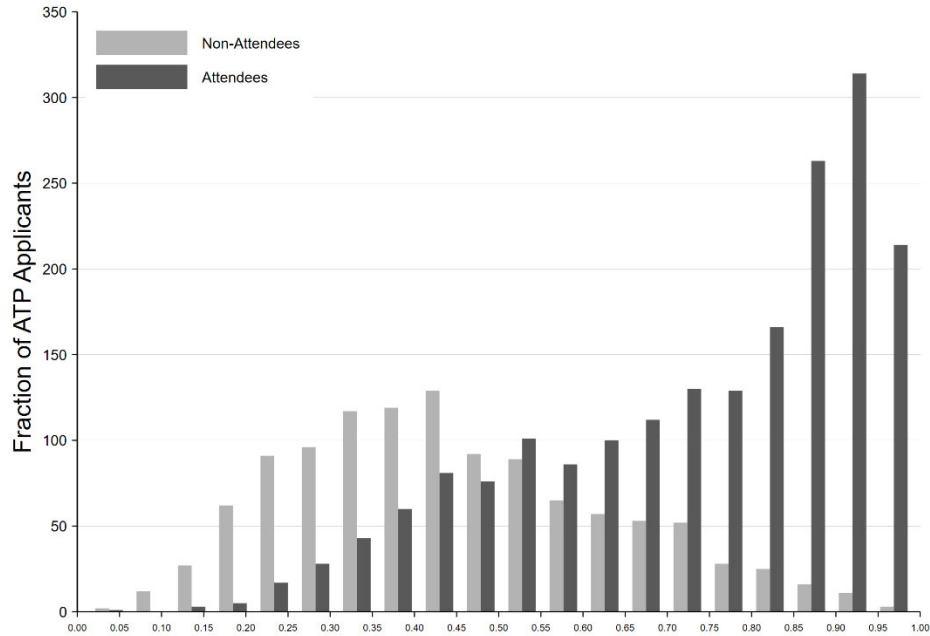
- Snyder, Thomas D. (Ed.). 1993. "120 Years of American Education: A Statistical Portrait." Washington, D.C.: U.S. Department of Education.
- Solow, Robert M. 1957. "Technical Change and the Aggregate Production Function." *The Review of Economics and Statistics*, 39(3): 312-320.
- Stinchcombe, Arthur L. 1965. "Social Structure and Organizations." In James G. March (Ed.), *Handbook of Organizations*, (pp. 142-193) Chicago, IL: Rand McNally.
- Svorenčik, 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	<i>t</i> -stat	Non-attendees	Attendees	<i>t</i> -stat
PhD	0.013 (0.003)	0.036 (0.004)	3.738	0.030 (0.011)	0.027 (0.003)	0.216
Age in the Year of Last Application	25.931 (0.042)	26.016 (0.033)	1.596	26.021 (0.056)	26.145 (0.139)	0.923
Applies more than once	0.027 (0.005)	0.028 (0.004)	0.154	0.043 (0.010)	0.028 (0.004)	1.434
Number of Applications	1.028 (0.005)	1.029 (0.004)	0.093	1.044 (0.010)	1.029 (0.004)	1.426
Number of Institutes Applied For	2.948 (0.061)	3.933 (0.053)	11.789	3.501 (0.115)	3.596 (0.057)	0.730
Number of Associate Tracks Applied For	1.828 (0.025)	2.068 (0.019)	7.811	1.962 (0.034)	1.991 (0.021)	0.729
AΩA Honor Medical Society	0.257 (0.013)	0.383 (0.011)	7.162	0.305 (0.019)	0.328 (0.016)	0.894
Pre-ATP Nb. of Publications	0.582 (0.037)	1.005 (0.039)	7.367	0.841 (0.080)	0.846 (0.045)	0.063
Pre-ATP JIF-weighted Nb. of Publications	3.288 (0.292)	6.595 (0.330)	6.835	6.107 (0.967)	5.289 (0.330)	0.854
NIH Grants for Applicant's Medical School	170.323 (3.792)	207.006 (3.438)	6.879	185.481 (6.033)	189.629 (7.441)	0.428
NIH Grants for Applicant's Internship Hospital	90.915 (2.590)	97.153 (1.882)	1.978	90.758 (3.505)	92.300 (3.863)	0.294
Attended Harvard Medical School	0.075 (0.008)	0.142 (0.008)	5.614	0.101 (0.014)	0.121 (0.008)	1.238
Attended Johns Hopkins School of Medicine	0.047 (0.006)	0.059 (0.005)	1.356	0.043 (0.007)	0.056 (0.006)	1.576
Attended Columbia University	0.050 (0.006)	0.044 (0.005)	0.725	0.051 (0.008)	0.042 (0.005)	0.965

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. Descriptive statistics: 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

	<i>Program Selection</i>			<i>Informative Censoring</i>		
	<i>Parsimonious</i>		<i>Saturated</i> Model [Lasso]	<i>Parsimonious</i>		<i>Saturated</i> Model [Lasso]
	Model [Logit]			Model [Logit]		
	(1a)	(1b)	(1c)	(2a)	(2b)	(2c)
Log(Pre-ATP Nb. of Publications)		0.307** (0.071)	0.329** (0.071)		-0.192** (0.064)	-0.210** (0.066)
Ln(NIH Grants for Applicant's Medical School)	0.357** (0.090)	0.317** (0.091)		-0.193** (0.067)	-0.158* (0.066)	
Ln(NIH Grants for Applicant's Internship Hospital)	0.019* (0.009)	0.017† (0.010)		-0.031** (0.008)	-0.029** (0.009)	
PhD	0.932** (0.334)	0.577† (0.342)	0.794* (0.311)	-1.347** (0.354)	-1.036** (0.358)	-1.141** (0.359)
No Internship	1.962* (0.839)	1.763* (0.849)	1.117 (0.981)	-2.716* (1.056)	-2.583* (1.068)	-3.936** (0.879)
Applies more than once	-0.039 (0.300)	-0.091 (0.295)	0.071 (0.273)	0.076 (0.246)	0.115 (0.249)	-0.028 (0.242)
AΩA Honor Medical Society	0.688** (0.105)	0.701** (0.106)	0.661** (0.101)	-0.346** (0.087)	-0.346** (0.088)	-0.347** (0.087)
Constant	-3.278† (1.748)	-2.649 (1.775)		3.049* (1.311)	2.329† (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 ([†] $p < 0.10$, * $p < 0.05$, ** $p < 0.01$). *Sources:* ATP Index Cards, PubMed, CGAF.

Table 3. Career choice outcomes

	X-Sect.	Lasso Weights	
	Naive	ATE	ATET
<i>Poisson Estimates</i>			
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)
<i>Logit Estimates</i>			
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 $100 \times (1.091 - 1) = 9.1\%$ 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. Naïve	Lasso Weights	
		ATE	ATET
Career Nb. of Pubs	1.922** (0.146)	1.639** (0.128)	1.668** (0.152)
Career Nb. of Pubs, First/Last Authorship Position	1.970** (0.146)	1.673** (0.131)	1.701** (0.157)
Career Citations	2.316** (0.227)	1.775** (0.190)	1.884** (0.220)
Nb. of Patents	2.515** (0.521)	1.635* (0.368)	1.559† (0.406)
Career Citations to Pubs in Patents	3.015** (0.511)	1.892** (0.386)	1.938** (0.454)
Career NIH Grants	2.558** (0.591)	1.846* (0.489)	1.807* (0.517)
Career NIH R01 Grants	2.285** (0.352)	1.668** (0.296)	1.774** (0.340)
Nb. NIH-R01-Funded Trainees	2.410** (0.365)	1.621* (0.349)	1.689* (0.429)
Trainee Career NIH R01 Grants	2.677** (0.501)	1.890** (0.423)	2.039** (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 $100 \times (1.922 - 1) = 92.2\%$ 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 (with citation data available)	1.951** (0.151)	1.637** (0.130)	1.680** (0.156)
Career Nb. of Pubs Top 50% of the Citation Distribution	2.065** (0.169)	1.696** (0.146)	1.735** (0.175)
Career Nb. of Pubs Top 25% of the Citation Distribution	2.157** (0.189)	1.717** (0.162)	1.771** (0.194)
Career Nb. of Pubs Top 5% of the Citation Distribution	2.348** (0.247)	1.800** (0.202)	1.884** (0.240)
Career Nb. of Pubs Top 1% of the Citation Distribution	2.653** (0.347)	1.963** (0.284)	2.176** (0.322)
Career Nb. of Pubs Top 0.1% of the Citation Distribution	2.814** (0.533)	1.968** (0.401)	2.195** (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 $100 \times (2.814 - 1) = 181.4\%$ 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. Naive	Lasso Weights	
		ATE	ATET
Career Nb. of Pubs	1.609** (0.117)	1.478** (0.112)	1.505** (0.131)
Basic Science Articles	2.787** (0.320)	2.197** (0.291)	2.184** (0.348)
Translational Medicine Articles	1.830** (0.196)	1.542** (0.179)	1.570** (0.207)
Clinical Trial Articles	1.584** (0.188)	1.544** (0.178)	1.674** (0.205)
Other Clinical Articles	1.056 (0.092)	1.121 (0.108)	1.132 (0.121)
Articles Appearing in Translational Journals	2.545** (0.424)	2.048** (0.357)	2.188** (0.457)
Inspires Translational Research	1.799** (0.211)	1.583** (0.186)	1.623** (0.207)
Builds on Translational Research	1.692** (0.213)	1.595** (0.197)	1.728** (0.224)
Articles Cited in Patents	2.138** (0.226)	1.767** (0.215)	1.800** (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 $100 \times (2.138 - 1) = 113.8\%$ 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 ([†] $p < 0.10$, * $p < 0.05$, ** $p < 0.01$). *Sources:* ATP Index Cards, PubMed.

Table 7a: Robustness analyses

	IPTC Lasso Weights			CEM	Zero Pre-ATP Pubs	
	No Winsoring	Winsoring, 95 th pctl.	Winsoring, 90 th pctl.		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 $100 \times (1.639 - 1) = 63.9\%$ 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 ([†] $p < 0.10$, * $p < 0.05$, ** $p < 0.01$).

Table 7b: Robustness analyses

	Nb. of Pubs		$\sinh^{-1}(\text{Nb. of Pubs})$	
	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 ([†] $p < 0.10$, * $p < 0.05$, ** $p < 0.01$).

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 postgraduate 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 and 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 the 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, shows 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.*

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 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

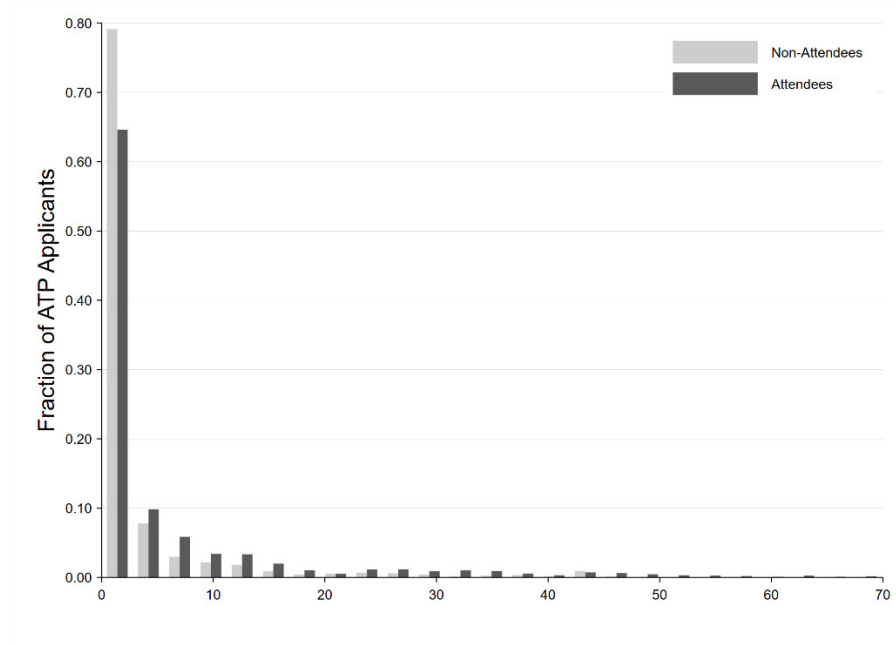
* In the subsamples for each track, we only include as controls trainees who applied unsuccessfully for that track.

often worked in the same labs, and the distinction between research and clinical time was not always clear-cut in practice.

References

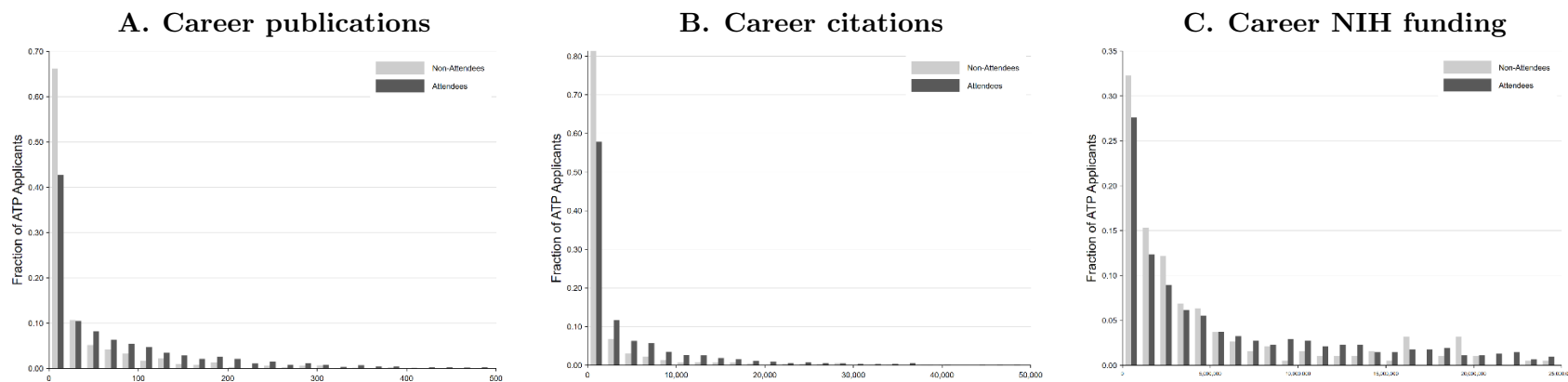
- Ellison, Glenn, and Edward L. Glaeser. 1997. "Geographic Concentration in U.S. Manufacturing Industries: A Dartboard Approach." *Journal of Political Economy*, 105(5):889-927.
- Nagle, Frank, and Florenta Teodoridis. 2019. "Jack of all trades and master of knowledge: The role of diversification in 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

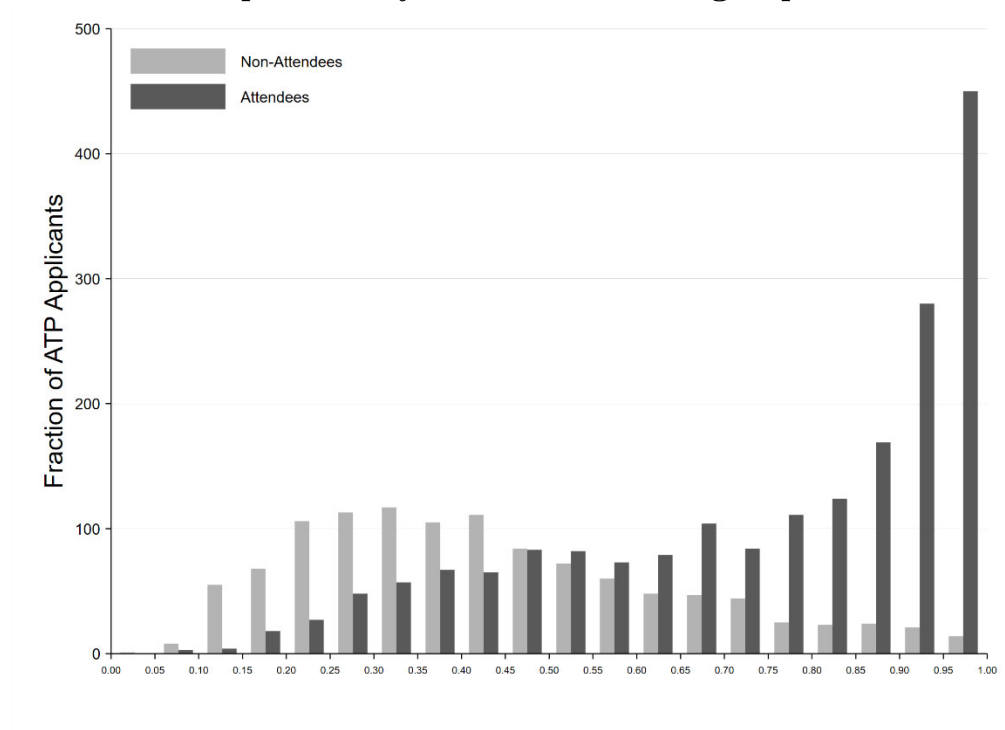


Note: Histogram for the number of original journal publications over the entire post-training career. Twenty-five outliers with more than 500 post-training publications omitted. 86% of applicants publish one article or more after career independence (91% of attendees; 78% of non-attendees). *Sources:*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

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.

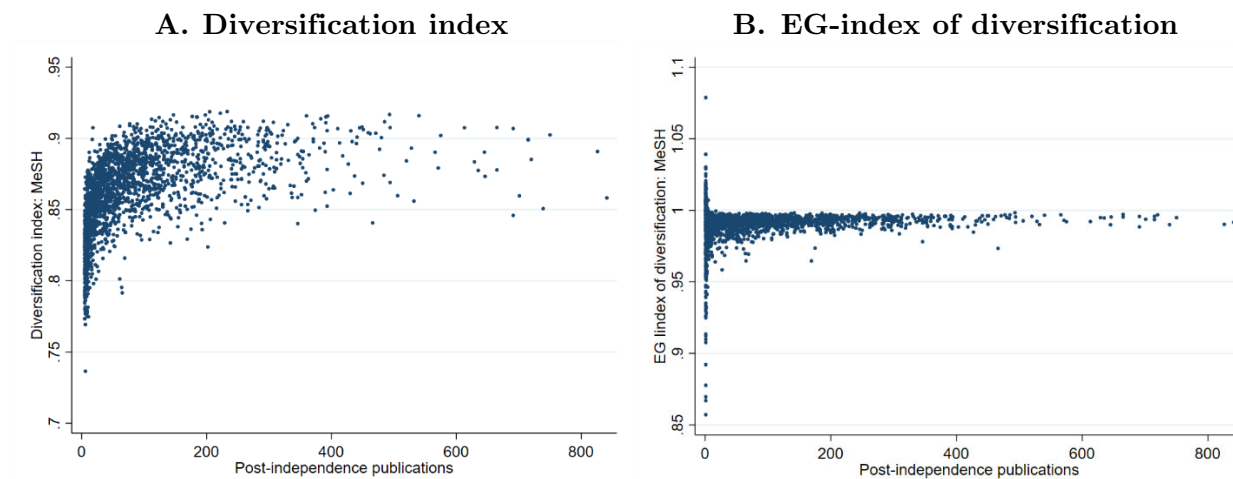
Note: Histogram for the cumulative NIH funding received over the entire post-training career (2015 dollars, deflated by the Biomedical R&D PPI). 1,087 fellows receive at least some NIH funding during their career. Two hundred and eighty four outliers with more than \$25 mln. in career funding omitted.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

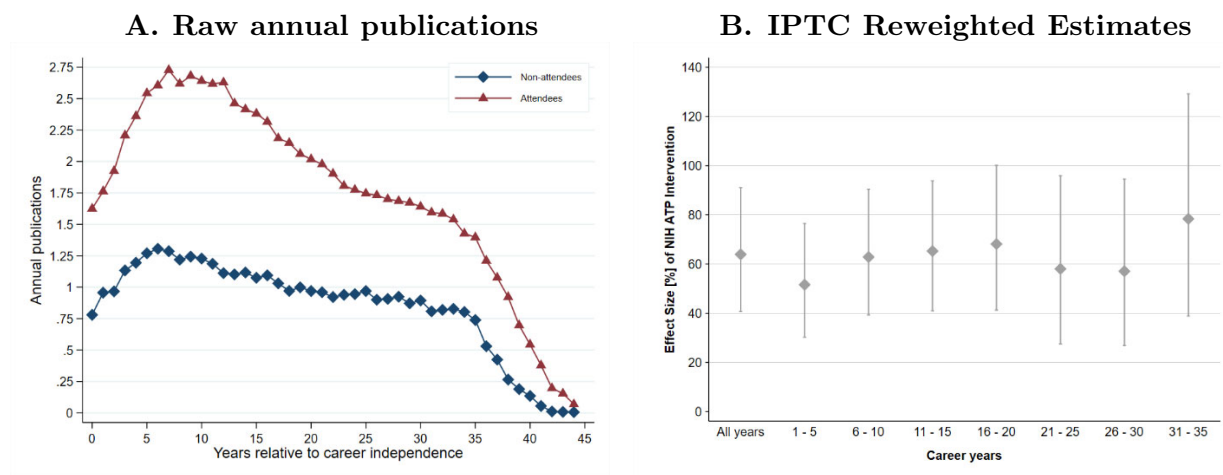
Figure B4. Specialization and career research productivity



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



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)
<i>N</i>	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)
<i>N</i>	3,075		

Note: Column percentages in parentheses.

Table B4a. Covariate balance using logit weights

	Unweighted sample			Logit IPTW Reweighting		
	Non-attendees	Attendees	<i>t</i> -stat	Non-attendees	Attendees	<i>t</i> -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. Naive	Lasso Weights		Logit Weights	
		ATE	ATET	ATE	ATET
<i>Logit Estimates</i>					
Researcher First Job	0.212** (0.018)	0.170** (0.022)	0.153** (0.026)	0.179** (0.026)	0.161** (0.033)
Ends Career as Researcher	0.216** (0.019)	0.192** (0.030)	0.175** (0.025)	0.140** (0.022)	0.175** (0.030)
<i>Poisson Estimates</i>					
Career Nb. of Pubs	1.922** (0.146)	1.639** (0.128)	1.668** (0.152)	1.698** (0.150)	1.752** (0.196)
Career citations	2.316** (0.227)	1.775** (0.190)	1.884** (0.220)	1.753** (0.217)	1.954** (0.293)
Nb. of Patents	2.515** (0.521)	1.635* (0.368)	1.559† (0.406)	1.252 (0.306)	1.290 (0.412)
Career NIH R01 Grants (\$ 2015)	2.285** (0.352)	1.668** (0.296)	1.774** (0.340)	1.442† (0.307)	1.713* (0.429)
Nb. NIH-R01-funded trainees	2.410** (0.365)	1.621* (0.349)	1.689* (0.429)	1.348 (0.396)	1.563 (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 $100 \times (1.922 - 1) = 92.2\%$ 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 († $p < 0.10$, * $p < 0.05$, ** $p < 0.01$). *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
<i>Logit Estimates</i>				
Researcher First Job	0.160** (0.018)	0.111** (0.021)	0.176** (0.019)	0.139** (0.023)
Ends Career as Researcher	0.146** (0.020)	0.131** (0.032)	0.188** (0.021)	0.155** (0.032)
<i>Poisson Estimates</i>				
Career Nb. of Pubs	1.922** (0.146)	1.639** (0.128)	1.739** (0.133)	1.483** (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 $100 \times (1.922 - 1) = 92.2\%$ 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 (0.438)	0.812 (0.450)	0.202* (0.148)	0.349 (0.324)	1.071 (0.152)	1.072 (0.138)	1.155 (0.172)	1.126 (0.136)
Two years of NIH training	1.486** (0.118)	1.415** (0.158)	1.158 (0.342)	1.200 (0.329)	1.099** (0.025)	1.163** (0.038)	1.165** (0.030)	1.239** (0.036)
Three years of NIH training	2.063** (0.211)	2.529** (0.322)	3.654** (1.492)	2.285** (0.550)	1.187** (0.037)	1.195** (0.043)	1.256** (0.040)	1.242** (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† (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 $100 \times (1.368 - 1) = 36.8\%$ 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 $100 \times (1.048 - 1) = 4.8\%$ 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. (i $p < 0.10$, $*$ $p < 0.05$, $**$ $p < 0.01$). *Sources:* ATP Index Cards, PubMed, Author-ity.

Table B9. Researcher specialization

	Diversification index		EG-index of diversification	
	Naive	Lasso Weights ATE	Naive	Lasso Weights ATE
ATP Attendee	0.0303** (0.0031)	0.0239** (0.0033)	0.0004 (0.0005)	0.0009 (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 (ⁱ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$).

Table B10. Research and training outcomes by weighting procedure

	Naive Estimates X-Sect.	Rewighted Estimates IPTW	Rewighted Estimates IPCW	Rewighted Estimates IPTCW
<i>Logit Estimates</i>				
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)
<i>Poisson Estimates</i>				
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 $100 \times (1.922 - 1) = 92.2\%$ 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
<i>Poisson estimates for full sample</i>								
Nb. of Pubs, Training Period	2.405** (0.142)	2.054** (0.163)	2.426** (0.179)	2.199** (0.168)	2.525** (0.161)	2.200** (0.167)	2.587** (0.278)	2.295** (0.263)
Number of Applicants	3,075	3,075	1,940	1,940	2,460	2,460	1,044	1,044
<i>Poisson estimates for research style sample</i>								
Nb. of Pubs, Training Period	1.516** (0.076)	1.432** (0.077)	1.520** (0.092)	1.507** (0.103)	1.562** (0.082)	1.495** (0.087)	1.572** (0.144)	1.450** (0.146)
Basic Science Articles	2.881** (0.253)	2.562** (0.242)	2.721** (0.308)	2.454** (0.310)	2.791** (0.271)	2.417** (0.261)	2.696** (0.451)	2.613** (0.471)
Translational Medicine Articles	1.834** (0.136)	1.670** (0.136)	1.820** (0.179)	1.756** (0.190)	2.041** (0.162)	1.853** (0.168)	1.719** (0.248)	1.561** (0.230)
Clinical Trial Articles	1.707** (0.341)	1.651* (0.331)	1.662* (0.375)	1.729* (0.376)	2.052** (0.374)	2.154** (0.417)	1.805 (0.657)	1.666 (0.569)
Other Clinical Articles	0.881† (0.065)	0.900 (0.079)	0.797* (0.070)	0.825* (0.073)	0.925 (0.073)	0.945 (0.077)	0.860 (0.109)	0.795 (0.127)
Inspires Translational Research	1.749** (0.179)	1.569** (0.173)	1.820** (0.236)	1.727** (0.233)	2.039** (0.222)	1.899** (0.231)	1.682** (0.332)	1.739** (0.318)
Builds on Translational Research	1.734** (0.353)	1.678* (0.348)	1.793* (0.441)	1.943** (0.475)	2.028** (0.404)	2.201** (0.475)	1.584 (0.602)	1.571 (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 $100 \times (2.405 - 1) = 140.5\%$ 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** (0.018)	0.170** (0.022)	0.239** (0.050)	0.175* (0.070)	0.219** (0.022)	0.173** (0.026)	0.173** (0.040)	0.135** (0.042)
Career Nb. of Pubs	1.922** (0.146)	1.639** (0.128)	2.335** (0.491)	2.029** (0.489)	2.103** (0.186)	1.767** (0.164)	1.309 (0.224)	1.148 (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 $100 \times (1.922 - 1) = 92.2\%$ 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
<i>Logit Estimates</i>								
Researcher First Job	0.212** (0.018)	0.170** (0.022)	0.195** (0.023)	0.148** (0.025)	0.193** (0.021)	0.142** (0.025)	0.225** (0.032)	0.175** (0.035)
Number of Applicants	3075	3075	1,940	1,940	2,460	2,460	1,044	1,044
<i>Poisson Estimates for full sample</i>								
Career Nb. of Pubs	1.922** (0.146)	1.639** (0.128)	1.918** (0.181)	1.511** (0.143)	1.917** (0.164)	1.637** (0.145)	1.926** (0.258)	1.593** (0.211)
Number of Applicants	3075	3075	1,940	1,940	2,460	2,460	1,044	1,044
<i>Poisson Estimates for research style sample</i>								
Career Nb. of Pubs	1.609** (0.117)	1.478** (0.112)	1.641** (0.148)	1.430** (0.131)	1.625** (0.132)	1.480** (0.127)	1.585** (0.202)	1.433** (0.179)
Basic Science Articles	2.787** (0.320)	2.197** (0.291)	2.732** (0.392)	2.449** (0.349)	2.730** (0.370)	2.319** (0.328)	2.443** (0.501)	2.112** (0.444)
Translational Medicine Articles	1.830** (0.196)	1.542** (0.179)	1.872** (0.251)	1.627** (0.209)	1.927** (0.241)	1.548** (0.217)	1.532* (0.302)	1.380 [†] (0.264)
Clinical Trial Articles	(0.188)	1.544** (0.178)	1.543** (0.251)	1.221 (0.234)	1.698** (0.229)	1.587** (0.222)	1.531 [†] (0.340)	1.295 (0.288)
Other Clinical Articles	1.056 (0.092)	1.121 (0.108)	1.016 (0.119)	0.918 (0.108)	1.075 (0.100)	1.119 (0.107)	1.201 (0.199)	1.128 (0.171)
Inspires Translational Research	1.799** (0.211)	1.583** (0.186)	1.875** (0.273)	1.634** (0.235)	1.992** (0.274)	1.620** (0.231)	1.514 [†] (0.340)	1.325 (0.286)
Builds on Translational Research	1.692** (0.213)	1.595** (0.197)	1.675** (0.270)	1.294 (0.256)	1.854** (0.264)	1.662** (0.246)	1.684* (0.353)	1.367 (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 $100 \times (1.922 - 1) = 92.2\%$ 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 Author-ity (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 effects of applying

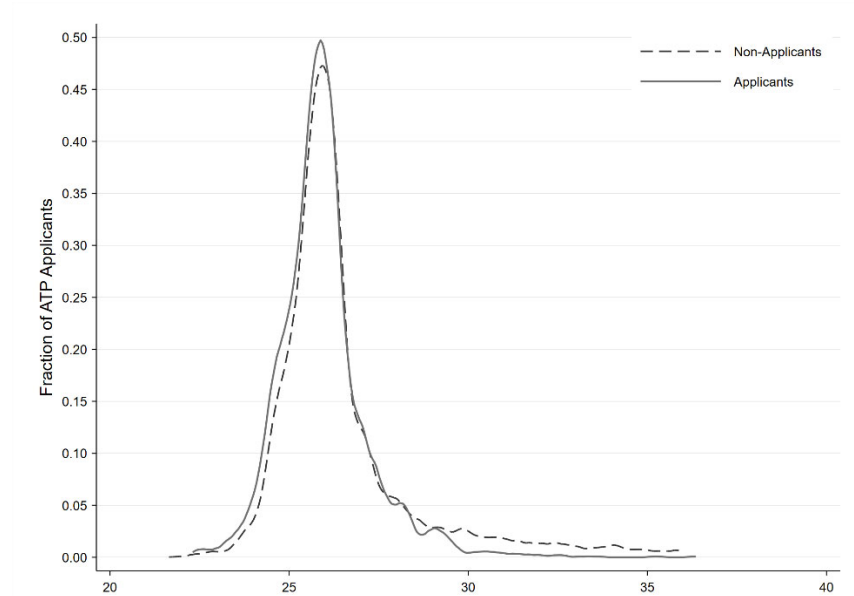
A natural question in light of the non-applicant sample is to understand the relative magnitude of selection into applying for the program with its treatment effect. Additionally, unsuccessful applicants may benefit from the exercise of applying itself (Ayoubi et al. 2019). Unfortunately, our research 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 ($\Omega\Omega\Lambda$), 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.
- Torrvik, 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 ($\times 1,000$)	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 ($\times 1,000$)	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 (0.103)	0.047 (0.100)
Log(Pre-ATP Nb. of Publications)	0.294** (0.078)	0.294** (0.078)	0.302** (0.079)
Ln(NIH Grants for Applicant's Medical School)	0.248* (0.109)	0.247* (0.109)	
Ln(NIH Grants for Applicant's Internship Hospital)	0.008 (0.011)	0.008 (0.011)	
PhD	0.744 (0.494)	0.747 (0.496)	1.087* (0.453)
No Internship			2.770* (1.143)
Applies more than once	0.119 (0.305)	0.117 (0.307)	0.433 (0.282)
AΩA Honor Medical Society	0.634** (0.121)	0.634** (0.121)	0.623** (0.115)
Constant	8.775** (2.178)	8.746** (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).²

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.³ 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).

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

² 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).

³ https://history.nih.gov/archives/oral_histories.html

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.⁴ 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.
- 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

⁴ 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.

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.⁵

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

$$E[y|X, Z] = \beta_0 + \beta'_1 Z + \beta_2 TREAT \quad (E.1)$$

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

$$\phi(X_i) = Prob(TREAT_i = 1|X_i) \quad (E.2)$$

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):

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

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

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

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.⁶ 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 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

⁵ 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.

⁶ 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).

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:

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

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:

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

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 φX174*, *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*, *arabidopsis thaliana*, *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.

Chapter 4

Does Grant Peer Review Penalize Scientific Risk Taking? Evidence from the NIH

Abstract

Does grant peer review punish risk taking? While risk is an inherent aspect of innovation, those projects with high degrees of risk may be more likely to lead to breakthrough innovation yet may face challenges in winning the support necessary to be carried out. To study this, I analyze 103,164 R01-equivalent grants from the National Institutes of Health, 1980-2015. To measure risk taking, I use four distinct approaches—extreme tail outcomes, disruptiveness, pivoting from an investigator’s prior work, and pursuing intellectually distant ideas from other investigators—that each capture a different aspect of what it means to take risks. After carefully controlling for investigator, grant, and institution characteristics, I measure the association between risk taking and grant renewal. Across each of these measures, I find grants with high levels of risk taking are renewed at lower rates than those with low levels of risk taking. The magnitudes of these effects are large: when comparing grants in the top and bottom deciles of risk taking, grants with greater risk taking have a 20.5%, 24.4%, 16.9%, and 12.4% decline in renewal rate for each measure of risk taking, respectively. I conclude with a discussion of the implications for policy-makers and managers of innovation for fostering risky research.

1. Introduction

Few topics elicit as much consensus as the idea peer review punishes risk taking. Reports abound, ranging from the popular press to U.S. Congressional testimony, from tweets to the pages of *Nature*, and are authored by NIH-funded investigators and leading science policy makers (Kent 2018, Kolata 2009, Kornberg 2007, Nielson 2022, Sutter 2022, Woolston 2014). So pervasive and prominent are these concerns that it leads scientists to write articles satirizing how risk-adverse grant funders are and muse if even former Nobel Prize winners could receive funding for their seminal discoveries in the modern funding climate (Fields 2014, Petsko 2012).

Yet, for all of the abundance of anecdote, there is surprisingly little evidence to support claims peer review punishes risk taking. Azoulay et al. (2011) is frequently viewed as evidence of this. It compares two grant funding mechanisms—the Howard Hughes Medical Institute Investigatorships and National Institutes of Health (NIH) R01 grants—and shows the former encourages more risk taking. This is different, however, from demonstrating that peer review directly punishes risk taking. Similarly, prior literature has examined how novelty shapes receiving grant funding, however, novelty is distinct from risk taking: investigators can pursue novel research which isn’t risky, and risky research which isn’t novel (Ayoubi et al. 2021, Packalen and Bhattacharya 2020, Veugelers et al. 2022). Beyond just theoretical interest, grant funders potentially punishing risk taking has very direct implications for what research is undertaken. In a recent survey of investigators supported by Fast Grants, 78% stated they would make significant changes to their research program if they were not constrained by the demands of grant funding agencies, including pursuing more ambitious research programs, pivoting to new topics, and testing hypotheses others see as unlikely to succeed (Collison et al. 2021).

More generally, the choice of what scientists choose to study is fundamental to determining the rate and direction of innovation. Not all proposed scientific projects, however, have an equal likelihood of extending the scientific frontier (Kuhn 1962). High risk, high reward projects may be more likely to lead to breakthrough innovation (Foster et al. 2015, Uzzi et al. 2013, Veugelers and Wang 2019, Wang et al. 2017). While both high risk, high reward projects and more incremental research are valuable, if the former are undersupplied, this may result in a slower rate of overall technological progress and ultimately lower economic growth (Rzhetsky et al. 2015). The costs of undersupplying risky research may be especially prominent for early-stage basic science research, which may have difficulty internalizing the value of its innovations and lead to substantial positive spillovers (Aghion et al. 2008, Azoulay et al. 2019, Fleming et al. 2019).

What causes individual scientists to select risky projects? Rather than being solely determined by an individual’s preference for risk taking, this choice is shaped in important ways by the incentives, organization environment, and supporting institutions that surround a researcher. Tolerance for failure (Tian and Wang 2014), long time horizons (Lerner and Wulf 2007), team structure (Lee et al. 2015, Wu et

al. 2017), and the structure of financial incentives (Graff Zivin and Lyons 2020) all shape how much risk innovators choose to bear when selecting projects. The choice of research project is made more challenging by the difficulty in predicting the ultimate benefits if the project is successful; this is especially the case for early stage and highly uncertain projects (Rosenberg 1995).

In this paper, I study how a specific institution, grant peer review, shapes scientific risk taking. In grant peer review, investigator scientific proposals are evaluated on their merit by a panel of experts, with the resulting scores used to determine grant funding. To study this, I need overcome three central challenges. First, I need a setting where I can systematically observe research grants and their associated outcomes. In particular, settings where secrecy plays a prominent role will make it difficult to measure the results of innovation projects. Second, I need the research for projects both with very high and very low risk to ultimately be undertaken. It will be difficult to estimate counterfactual outcomes if any punishment for risk taking prevents the project from ultimately being completed. Finally, I need a valid approach to measuring risk taking. Importantly, as investigators may have different levels of risk taking across different projects and their approach to risk taking may evolve over time, it is essential measures of risk taking are specific to the particular proposed project rather than reflecting an investigator’s historical risk taking.

I argue the NIH grant peer review process provides a setting that allows us to address these challenges. Funded NIH grants are in the public domain. In particular, strong norms in biomedicine lead most innovation to be published, providing a robust paper trail I can characterize (Dasgupta and David 1994). As NIH grants can be competitively renewed through a peer review process, focusing on risk taking in awarded grants lets us observe any punishment for risk taking for a set of research all of which was undertaken. Finally, I capitalize on the requirement of papers, as a condition of funding, to acknowledge their NIH grant support. This allows us to use rich bibliometrics in biomedicine to measure risk taking among research emerging directly from the grant.

I study how risk taking is penalized or rewarded in 103,164 NIH-funded R01-equivalent grant cycles, 1980-2015. After controlling for a robust set of covariates for investigator demographics and quality, institution quality, and grant characteristics, I observe the relationship between risk taking in publications that acknowledge funding from the grant and whether the grant was competitively renewed. Risk taking, much like creativity, is a nebulous concept. There is no consensus on what constitutes risk in science nor how it should be measured (Althaus 2005, Franzoni and Stephan 2023). As there are different aspects of risk taking, each of which may have distinct dynamics, it is unlikely a single measure of risk taking can fully represent the phenomenon. Instead, I develop four separate measures of risk taking, each of which captures a different way investigators may take risks, and examine if my findings depend on the particular measure used. First, risk taking may be associated with more extreme tail outcomes, with greater variance in outcomes and both more exceptionally good and exceptionally bad results (Azoulay et al. 2011). Second, risky research may be more disruptive, seeking to overturn the status quo, rather than consolidative, reinforcing earlier results (Funk and Owen-Smith 2017). Third, researchers may pivot from what they have

done in the past to follow new scientific opportunities, capturing the idea that risk isn't only inherent to the project itself but depends also on who is undertaking the project. Finally, researchers may stand out from the crowd, selecting projects that are intellectually distant from those of other investigators (Uzzi et al. 2013). Across each of these measures, I take a non-parametric approach in exploring effects across the full distribution of different levels of risk taking. This lets the data speak for itself in whether any observed effect is driven primarily by differences in conventional levels of risk taking, or alternatively, by extreme tail risk taking, arguably those projects with the most potential to lead to breakthrough innovation.

Across each of these measures, I find that risk taking is penalized. When comparing grant cycles in the top and bottom decile of risk taking, grants with greater risk taking have a 9.5% lower renewal rate (20.5% decline) when measuring risk taking using extreme tail outcomes, an 11.1% lower renewal rate (24.4% decline) when measuring risk taking by its disruptiveness, a 7.7% lower renewal rate (16.9% decline) when measuring risk taking by an investigator pivoting from her prior research, and a 6.1% lower renewal rate (12.4% decline) when measuring risk taking by standing out from what other investigators are studying. In contrast to my measures of risk taking, novelty is associated with higher grant renewal rates. This emphasizes how risk taking and novelty may have distinct dynamics, and that novelty is not always a valid proxy for risk taking in some settings. I also find the decrease in grant renewal for risk taking is larger for novel research, suggesting this penalty for risk taking results in a welfare loss. Finally, I examine heterogeneity in the penalty for risk taking by career age, gender, status, and grantsmanship. I show risk taking is penalized across each measure of risk taking for all subgroups examined. Additionally, I find suggestive evidence this penalty is larger for early career investigators.

This study's primary contribution is to provide direct evidence that grant peer review may punish scientific risk taking. In addition to the potential policy importance for public and philanthropic scientific grant programs, this emphasizes how the type of innovations a system produces is shaped—deliberately or not—by the institutions and managerial processes that support it. While I focus on peer review for scientific grants, this study has implications for the design of organizational processes to select innovative projects in other settings that value risk taking.

The rest of the paper proceeds as follows. In Section II I develop how peer review might shape scientific risk taking. Section III presents the setting, data, empirical strategy, and descriptive statistics. I report results in Section IV. Section V concludes with the implications of this study.

2. Peer review and risk taking

Peer review plays a prominent role in the selection of scientific projects (Chubin and Hackett 1990, Stephan 2012). While some degree of risk is inherent in nearly all innovation projects, a central challenge facing projects with high levels of risk is both that they are less likely to succeed, and that conditional on success, it may be more difficult to predict their ultimate impact and benefits (Rosenberg 1995). A number

of distinct processes may potentially be at play in shaping how peer review may punish risk taking (Azoulay and Li 2022, Franzoni et al. 2022). These may occur at the level of the investigator, individual peer reviewer, peer review committee, and funding institution. While some of these factors may be idiosyncratic to particular grant funding institutions, others may be present in how innovation projects are selected across a range of different settings.

Peer review may shape the degree of risk taking investigators pursue through several indirect mechanisms. Many academic labs rely on external grants to fund their operations (Stephan 2012). An inability to win external funding may result in closure of the lab, laying off dependent lab personnel, and loss of the investigator’s job and potentially career (Ruben 2017). To the extent investigators *perceive* grant peer review to punish risk taking, this will lead them to instead propose safer, more incremental projects (Kent 2018, Langer 2012, Nielson 2022). Similarly, scientific trainees may become imprinted with a less risky approach to research after observing cautious, incremental projects proposed by their mentors in response to a perceived penalty for risk taking (Azoulay et al. 2021, Higgins 2005). In this way, peer review may shape the supply of risk taking in research outside of its direct role in evaluating particular projects.

Penalization of risk taking may also occur at the level of individual peer reviewers. Individuals may have a psychological bias against the uncertainty that characterizes risky research compared to more incremental projects (Carson et al. 2022, Fox and Tversky 1995, Mueller et al. 2012). In line with this, evaluators of research proposals have been found to give lower scores to proposals with greater novelty (Boudreau et al. 2016) and to emphasize feasibility (Kreiger et al. 2022, Lane et al. 2022a). The bias against novelty may also lead to delayed recognition of the underlying potential of a scientific work (Wang et al. 2017).

Peer review is conventionally undertaken in a committee. Bias against risk taking can occur due to both committee processes and committee membership. Proposals with greater risk taking may have greater variance in reviewer assessment of their quality (Langer 2012). Committee processes that emphasize consensus and uniformly high evaluations across its members to be funded may penalize risk taking (Azoulay and Li 2022, Franzoni et al. 2022).¹ Similarly, given the wider variation in evaluations of risky projects, processes where each member independently fully evaluates a proposal prior to the committee meeting may be more likely to result in identifying a judge who enthusiastically supports the project compared to processes where one or two committee members have primary responsibility for evaluating each proposal and then present their recommendation to the committee. Exposure to negative assessments of other committee members prior to the final judgment may lead evaluators to revise their own assessment downwards more so than exposure to positive assessments leads them to revise their assessments upwards, biasing against projects with more disagreement (Lane et al. 2022b). Who serves on the committee is also

¹ For instance, a process where a proposal’s overall score is simply the average of each judge’s evaluation will emphasize consensus more than having scores aggregated after dropping the lowest judge score(s) or a process which emphasizes having a few judges with very high levels of support.

important. The penalty for novelty may be larger from those investigators in the same field as the novel proposal (Wang et al. 2017). This suggests intellectual diversity in committee membership may reduce the bias against risk taking. Additionally, a scientist’s prior risk taking may shape their likelihood of penalizing risk taking when evaluating project proposals. To the extent this is true, considering prior risk taking when determining committee membership may reduce the committee’s risk penalty (Nicholas and Ioannidis 2012).

Finally, some aspects of the potential for peer review to penalize risk taking emerge at the level of the funder and innovation ecosystem. Of course, funders may shape any potential risk-taking penalty by how they structure committee procedures and membership, set expectations for project stage and any required preliminary supporting evidence, and determine explicit evaluation criteria. Political pressure to show constituents funders are effective stewards of money may shift towards the selection of reliable, incremental research projects (Lorsch 2015). Risk taking may also be facilitated by shifting the focus of selection from individual projects to a *portfolio* of projects (Azoulay and Li 2022, Franzoni et al. 2022). Finally, the vigorous competition that emerges from the low overall rate of funding may push investigators and reviewers to prioritize short-term advances and inhibit creativity (Alberts et al. 2014, Fang and Casadevall 2015).

Prior literature has studied the association between aspects of research that are related to risk taking and the evaluation and funding of scientific proposals and publications. This includes novelty (Ayoubi et al. 2021, Boudraeu et al. 2016, Packalen and Bhattacharya 2020, Teplitskiy et al. 2022, Veugelers et al. 2022), interdisciplinary research (Banal-Estanol et al 2019, Bromham et al 2016), and the combination of novelty and feasibility (Krieger et al. 2022, Lane et al. 2022a). This work builds on these studies in several notable ways. First, I develop a set of eclectic measures of risk taking, each of which captures a different aspect of risk, and empirically emphasize effects across the full distribution of different levels of risk taking. Prior research has often conflated novelty with risk taking; I discuss the distinction between the two in more detail in section III and empirically distinguish between risk taking and novelty in my results. Second, I focus on measures of risk taking in the grant proposal itself and publications emerging directly from the grant. In contrast, prior studies have frequently focused on risk-related measures of an investigator’s research in the years preceding a grant application and used this to predict grant award. Yet, investigators not only evolve over time in their approach to risk taking, they frequently manage a portfolio of research projects each of which may have different levels of risk taking. An investigator’s past research may not accurately reflect risk taking in a focal grant proposal, and indeed, investigators may deliberately avoid proposing projects with high risk taking if they perceive the funding agency to punish this, instead either forgoing these projects or funding them through other sources. Boudraeu et al. (2016) experimentally assigns grant proposals and focuses on individual judge evaluations, while Carson et al. (2022) experimentally adjusts research project evaluations and focuses on individual funder project selection. These approaches provide important causal insights, but do not capture selection committee dynamics, realized investigator behavior after a grant is funded, as well as faces the usual challenges of generalizing from experimental to

real-world conditions. Finally, by studying the universe R01-equivalent NIH grants, I focus on a population of substantial policy interest.

3. Empirical design, data and descriptive statistics

Setting: The National Institutes of Health

The NIH plays a prominent role in funding biomedical research. In 2012, it comprised 27% of all biomedical research funding in the United States and nearly two-thirds of all public and philanthropic research funding (Moses et al. 2015). In 2021, the NIH spent \$32 billion to support extramural research projects (Lauer 2022). While the NIH funds clinical research (including clinical trials, health services research, and behavioral studies among others), the bulk of their funding supports basic research aimed at advancing the frontier of scientific understanding. With research project grants having an average size of over \$580,000 in 2021, NIH grants are a major source of funding for the researchers they support (Stephan 2012). Not only does this funding play an important role in funding lab materials and the salaries of lab employees, many academic researchers are in “soft money” positions, funding their own salary out of awarded grants. The ability to successfully raise grant funding is an important metric upon which academic employment and promotion decisions are made.

Understanding the process the NIH uses to evaluate and fund grant applications can provide useful context for my empirical analysis (Azoulay and Li 2022, Gerin et al. 2018). Briefly, the NIH is comprised of 27 separate institutes or centers (IC) which are typically organized around a set of diseases (e.g., National Cancer Institute) or organ systems (e.g., National Heart, Lung and Blood Institute). Each IC is responsible for funding research that is potentially relevant to its mission. Scientific evaluation of grant proposals occurs primarily in standing review committees known as study sections (e.g., “Cellular and Molecular Immunology” or “Atherosclerosis and Vascular Inflammation”). Each study section may review and score grants from multiple ICs, which are then funded by the ICs in order of their study section score until the IC exhausts its funds.² The amount of resources devoted to reviewing grant proposals—both from the NIH with staff to manage the process and from study section members—is substantial. In 2021, the NIH reviewed nearly thirty-eight thousand applications for R01-equivalent grants (Lauer 2022).

The NIH is an attractive setting for studying how peer review might respond to risk taking. First, professional norms lead most academic research to be published in peer-reviewed journals (Dasgupta and David 1994). When combined with rich bibliometrics in biomedicine, this provides a robust “paper trail” I can follow and characterize to measure the degree of grant risk taking. Second, as NIH research is a source of significant improvements in health and positive economic spillovers, understanding how the NIH may

² Institute directors do have discretion to fund applicants out of order considering factors such as fit with the institute’s mission, the overall portfolio of projects being supported, and their own evaluation of application quality.

more effectively utilize its funds to support breakthrough innovation is of considerable policy interest (Azoulay et al 2019, Fleming et al. 2019).

Data

As I do not have observe grant applications, I focus empirically on whether an awarded grant was renewed. I rely on the NIH's Consolidated Grant Applicant File to identify all those researchers who received an NIH grant, 1980-2015. I start the analysis in 1980 as I am not able to reliably link publications to the NIH grants which supported them before this date, and end in 2015 to allow sufficient time for follow up to observe research outcomes and renewal status of grants awarded in this year. I limit the analysis to R01-equivalent grants.³ These investigator-initiated grants are a major funding mechanism to support the work of principal investigators who have achieved career independence and run their own laboratory. I also exclude grant awards from special emphasis study sections, which may not have an option for grant renewal, and grants without associated publications for which many of my measures of risk taking would be undefined. I focus the analysis on the grant cycle, which corresponds to a single competitive review of a funding proposal. A cycle starts the year a grant proposal is funded and ends either at the conclusion of grant funding or when a proposal for grant renewal is competitively reviewed and funded. As grants may have multiple principal investigators and investigator characteristics are explicitly considered during renewal decisions, I analyze the sample at the investigator-grant-cycle level.⁴ This gives a final sample of 103,164 investigator-grant-cycles from 63,101 grants and 37,222 investigators.

Publication data is from *PubMed*, an online resource from the National Library of Medicine to support access to the biomedical research literature that indexes over 30,000 journals and over 34 million articles. I limit the analysis to original research publications; this excludes other types of publications such as review articles, letters, and erratum. For each grantee, I identify their career publications in *PubMed* using Author-ity (Torvik and Smalheiser 2021, Torvik and Smalheiser 2009, Torvik et al. 2005). Author-ity disambiguates authors using the insight that authors of the same name sharing coauthors, affiliations, keywords, and other such characteristics are likely, in fact, to be the same person. Author-ity has been shown to be highly accurate (Lerchenmueller and Sorenson 2016).

I use NIH RePORTER to link publications attributable to each NIH grant. This data relies on the grant having been listed in the publication's acknowledgement section as a funding source, which is required as a condition of accepting NIH funding. I consider publications attributable to a grant cycle if they are

³ I consider R01-equivalent grants to be those with an NIH activity code of R01 or R37. I exclude several grant mechanisms considered by the NIH to be R01-equivalent but which may have distinct dynamics around risk taking. For instance, I exclude DP1, DP2 and DP5 which are part of the NIH's effort to fund high risk, high reward projects; R56 and RL1 which may not be applied for by the investigator; U01 in which NIH staff actively support the research; and R35 and RF1 which provide longer-term support to investigators.

⁴ One hundred thirty eight grant cycles appear in the sample more than once as they have multiple principal investigators.

published up to one year after the last year of the grant cycle to allow for publication delays.⁵ As investigators may have different degrees of risk taking across different projects and funding sources, in all cases, my measures of risk taking are calculated solely using publications during a grant cycle that acknowledge funding from the grant rather than also including publications during the grant cycle which do not acknowledge funding from the grant.

To measure research topic and methodology, I capitalize on Medical Subject Headings (MeSH). MeSH is a hierarchical controlled vocabulary created and maintained by the National Library of Medicine used for indexing and searching the biomedical literature. In the 2019 edition of MeSH, there are over 29,000 terms, with an average of about 12 MeSH keywords associated with each PubMed publication. Crucially, MeSH are assigned by professional indexers. This excludes any concern authors selecting MeSH might do so strategically (Bachrach and Charen 1978). While it is difficult to assess accuracy as there is no single “correct” way to index articles, MeSH have generally been found to be consistent across indexers (Coletti and Bleich 2001, Funk and Reid 1983). In addition to MeSH associated with grant publications, I use the National Library of Medicine Medical Text Indexer to assign MeSH to grant abstracts. Grant abstract text is available from the NIH RePORTER.

I supplement publication data with citations from Thomas Reuter’s *Web of Science*. All grants are adjusted to 2020 dollars using the Biomedical Research and Development Price Index.

Measuring risk taking

There is not consensus on what constitutes risk in science nor how it should be measured (Althaus 2005, Franzoni and Stephan 2023). For instance, within science, there are different types of risk, including technical risk stemming from what nature’s answer is to the question asked, execution risk in being able to successfully carry out the planned experiments, competitive risk from other investigators studying the same or similar question, and risk in how a successfully completed project will be perceived by peers or advance the investigator’s career. Some aspects of scientific risk are not inherent to the science itself, but also depend on who is undertaking the study and the environment that supports it. My measures of risk taking do not correspond one-to-one with these different types of risk, but instead emphasize different combinations of them.

To illustrate how projects may have different aspects of risk, consider the challenges of a single measure in fully capturing risk taking dynamics across disparate projects. The Human Genome Project, once well underway, had little technical risk or danger of being viewed as a marginal contribution, but substantial competitive risk in the race to be the first to sequence the genome. In contrast, early efforts to

⁵ Among publications acknowledging a focal NIH grant in the sample, 20.8% are published more than one year after the end of grant funding. These publications and the results they are based on will typically not have been available to study sections when they review a grant for competitive renewal and so do not inform the outcome or measures of interest.

develop cancer immunotherapy had very substantial technical risk, as it may be very difficult or impossible to develop immunotherapy specifically targeting cancer but not normal tissue, but much lower levels of competitive risk. Similarly, studying new systems to encourage handwashing may have little technical or competitive risk, but may entail career risk from being viewed as a marginal contribution.

Additionally, connecting scientific novelty with risk taking is particularly challenging (Ayoubi et al. 2021, Foster et al. 2015, Uzzi et al. 2013). While pursuing novel research may be linked in some cases with risk taking, it conflates technical risk, competitive risk, and career risk in ways that may vary across different projects. More importantly, investigators can take risks without being novel, and can be novel without taking large risks. Consider efforts to develop non-mRNA COVID-19 vaccines. While this research was novel, targeting a new disease, there was arguably only modest levels of technical risk. In line with this, many research groups around the world were independently successful in developing such a vaccine, and some experts viewed it is a matter of when, not if, such efforts would be successful despite its novelty. In contrast, consider Judah Folkman’s initial findings on the role of angiogenesis in cancer. Other scientists viewed building on this work as so technically risky few were willing to do so until over a decade later, by which time arguably such follow on work may no longer be considered novel.

As an additional complicating factor in taking novelty as a proxy for risk taking, scientists are judged for employment and promotion on their perceived scientific contribution. If a scientist persistently undertakes work which isn’t novel or cutting edge, the resulting career risk of being viewed as having only low-impact contributions may be greater than the technical risk of studying cutting edge science (Franzoni and Stephan 2023). Additionally, even if a funding agency is risk averse, it still has a mandate to pursue new-to-the-world frontier science, and thus may look for “incremental or safe novelty” in projects. I see the study of factors which promote scientific novelty as essential research and novelty as a valuable metric in and of itself. Yet, I believe care should be taken in considering novelty as a proxy for risk taking.

To address the challenges of measuring risk taking, I employ four separate approaches. I argue each of these measures captures a distinct aspect of scientific risk taking. As each measure has different strengths and weaknesses in how they correspond to risk taking, by integrating across them I can develop a more complete understanding than examining each separately. In each case, I am non-parametric in my approach and examine the full distribution of different levels of risk taking. This lets the data speak for itself for whether any risk penalty is due primarily to differences in more conventional levels of risk taking or is instead driven by risk taking in the tails of the distribution.

First, risk taking may be associated with both more exceptionally good and exceptionally bad outcomes (Azoulay et al. 2011). Projects with this type of risk reflect a higher reward if successful, but also a higher chance of failure. In line with this, work with atypical combinations of ideas has been found to have a wider dispersion in outcomes (Schilling and Green 2011, Wang et al. 2017). I term this measure “extreme tail outcomes” and operationalize it as the difference between the highest and lowest vintage-adjusted citation percentiles among grant cycle publications. I focus on long-run vintage-adjusted citation

percentiles here for several reasons. Early citations can be noisy and there may be a delay in recognizing the contributions of highly novel articles (Wang et al. 2017). This lets us focus on the long-run view of the publication’s underlying quality and contribution. Additionally, this allows us to readily compare articles across different publication years.

Research may either be consolidative, reinforcing earlier results, or disruptive, overturning the status quo. Another way of taking risks is to undertake this type of disruptive research. To provide evidence of this, I employ the disruption index proposed by Funk and Owen-Smith (2017) and previously used in the study of science (Bornmann et al. 2020, Wu et al., 2019). In this index, a highly disruptive article is one in which future research cites the focal paper but not the focal paper’s references, while highly consolidative research has future research citing both the focal paper and the focal paper’s references. My measure of “disruption” for a grant cycle is the maximum of the disruption index percentile, calculated relative to all publications in that year in PubMed, among grant cycle publications.

As a third measure, I focus on intellectual distance between grant cycle publications and an investigator’s prior work. This takes seriously the idea that some aspects of risk are not inherent to the project itself but are also a function of who is undertaking the research. Scientists recombining their existing expertise to address new topics, employ new techniques, or explore newly identified opportunities may lead to promising previously unexplored findings. On the other hand, such exploration may devalue their prior expertise and decrease the quality of their output after pivoting (Arts and Fleming 2018, Hill et al. 2021).⁶ I term this type of risk taking “pivoting.” To capture this, I identify the fraction of MeSH term pairs among grant cycle publications that are new compared to the MeSH pairs from an investigator’s publications in the 5 years preceding the grant cycle.

As a final approach to measuring risk taking, I examine intellectual distance from what other NIH-funded investigators are studying. This reflects scientists studying unique recombinations of ideas and techniques (Fleming 2001, Uzzi et al. 2013). This speaks to the risks of being contrarian by choosing projects that do not fall within the mainstream of what their peers are studying and emphasizes career and competitive risks (Fang and Casadevall 2015). I refer to this measure as “standing out” from the crowd and measure it by identifying the fraction of MeSH term pairs from grant cycle publications in which the focal investigator is the only researcher using this term pair among all funded grant proposal abstracts in the same IC-study section-year.^{7,8}

⁶ Franzoni and Stephan (2023) point out that persisting in a researcher’s current trajectory rather than pivoting isn’t always associated with lower risk. If the current research line shows little promise, there may be less risk in pivoting.

⁷ The relevant sample size varies slightly across each of these measures. Extreme tail outcomes is only defined for those grant cycles resulting in at least two publications indexed in Web of Science. Disruption is only defined for those grant cycles resulting in at least one publication that is cited at least once. Pivoting is defined only if there is at least one publication with MeSH terms in both the current grant cycle and the preceding 5 years. Standing out is defined only if there is one publication with MeSH terms in the current cycle and funded grant proposal abstracts in the same IC-study section-year are available.

⁸ I compare an investigator’s grant cycle publication MeSH to that for all available awarded grant proposal abstracts, including those that are not R01-equivalent grants, in the same IC-study section-year. Among the 103,164

I also explore novelty to empirically distinguish it from risk taking. Prior literature has used a number of different approaches to measuring novelty, including examining unusual or intellectually distant combinations of references (Shibayama et al. 2021, Uzzi et al. 2013, Wang et al. 2017), unusual combinations of patent classes (Fleming 2001, Verhoeven et al. 2016), citation network centrality (Shibayama and Wang 2020), keyword and keyword combination age (Misha and Torvik 2016, Packalen and Bhattacharya 2020), and phenomena-specific measures such those based on chemical structure (Krieger et al. 2022, Rzhetsky et al. 2015). I follow Boudreau et al. (2016) in focusing on new combinations of MeSH terms. My measure of novelty is the fraction of MeSH pairs among grant cycle publications that were first used by any publication within PubMed within the past three years.⁹

Empirical strategy

I observe the universe of NIH grants actually awarded, but unfortunately, do not observe grant applications which are not ultimately funded by the NIH. Because of this, I empirically focus on NIH R01-equivalent grants that were awarded and measure whether or not the grant was renewed. This approach lets us reliably observe all investigators who were at risk of having their grant renewed. Among this population, a grant may not be renewed either because the investigator applied for but did not receive a competitive renewal, or because they never applied for a renewal. Given the career importance of grant funding in maintaining a lab, I expect most investigators will apply for renewal.¹⁰ Furthermore, if an investigator with risky research is dissuaded from applying for grant renewal due to their perception of a low likelihood of renewal and the time involved in applying, the implications of this for my research question are very similar to having applied and not been selected. As my sample evaluates the degree of risk taking among scientists all of whom have an R01-equivalent grant, this suggests these scientists will have similarly benefited from the act of applying for and receiving an R01-equivalent grant (Ayoubi et al 2019, Jacob and Lefgren 2011).

The largest threat to unbiased estimation of the impact of risk taking on grant renewal is potential omitted variable bias: that there is a factor correlated with both risk taking and grant renewal that is not included in my model. To address this, I use a rich set of controls for investigator, institution, and grant cycle observable characteristics.¹¹

investigator-grant cycles in the sample, I have abstracts for other grants in the same IC-study section-year for 83,451. The mean number of abstracts in the comparison group for each investigator-grant-cycle is 99. I prefer this approach to alternatives, as while data is not available for all possible investigator-grant-cycles, the comparison group is thicker for those that are available, facilitating my focus on the tail-measures of intellectual distance from other investigators.
⁹ 7.9% of MeSH pairs in the sample have three or fewer years between when they were first used in PubMed and their use in a grant cycle publication. My findings are robust to using thresholds other than three years, including using MeSH pairs new to PubMed in the past year.

¹⁰ Applications for competitive grant renewal, using the NIH's more inclusive definition of R01-equivalent (see footnote 3), were funded at a rate of 34.5% in 2015 (Lauer 2016). This is similar to the 35.9% renewal rate in my sample in the same year, consistent with most grantees applying for renewal.

¹¹ The NIH instructs reviewers to score each application based on five criteria: significance, innovation, approach, investigator, and environment. While the first three assess the importance, novelty, and feasibility of the proposed

I also consider controls related to grant research outputs. Study sections reviewing applications for grant renewal will consider an investigator's progress and achievement during the current cycle as reflected in its publications. Omitting measures of research productivity and impact during a grant cycle might lead to omitted variable bias. For instance, there may be some element of investigator quality that is correlated with both research productivity and risk taking during a grant cycle. As the association between risk taking and research quantity and quality may vary in magnitude and direction across different levels of risk taking, it is difficult to predict the direction of any bias if these controls are excluded. On the other hand, publication quantity and quality is a *result* of the mechanism through which risk taking may impact grant renewal (Angrist and Pischke 2008). Due to this, their inclusion may also result in biased estimates of the effect of risk taking on grant renewal. For this reason, I present each core result both with and without controls for publication quantity and quality during the grant cycle and view my estimates as an association rather than causal effect.

Econometric considerations

My primary estimating equation relates the dependent variable of interest, $renewal_{ij}$, of investigator i and grant cycle j . Formally, I estimate variations of

$$E[renewal_{ij} = 1 \mid \delta_i, \gamma_j] = \beta_0 + \beta_1 risktaking_{ij} + \delta_i + \gamma_j$$

where $renewal_{ij}$ is an indicator variable equal to 1 if the grant cycle was renewed and 0 otherwise. I include high dimensional vectors of controls for investigator and grant cycle characteristics, δ_i and γ_j , respectively. This includes sets of indicator variables for investigator age, degree and gender; log investigator career citations, log prior investigator NIH funding, an indicator variable for no prior investigator NIH funding, and log institutional NIH total funding as measures of investigator and institutional quality; and sets of indicator variables for year of evaluation, grant length, and the interaction between the IC and study section. Select specifications also include controls based on measures of research quality and quantity during the grant cycle, including log of mean journal impact factor, sets of indicator variables for the number of publications during the grant cycle separately both for those publications acknowledging and not acknowledging funding from the grant, and indicator variables for having publications in highly selective journals.^{12,13} I use robust standard errors with two-way clustering at the investigator and grant throughout. The maintained assumption is that, after conditioning on observables, measures of risk taking are uncorrelated with the error term.

project itself, the last two evaluate investigator qualifications and ability of the institution to support the work, respectively.

¹² I use journal impact factor as a proxy for publication quality for grant cycle publications as this metric is visible to peer review committees at the time of evaluation and early citations are a noisy signal of quality. My results are robust to instead controlling for citations at the time of grant cycle evaluation and long-run citations.

¹³ For this measure I include those publications which were published in *Nature*, *Cell*, *Science*, and the *New England Journal of Medicine*.

My primary specification is a linear probability model (LPM). Since LPM may generate fitted values outside of the interval $[0,1]$, its use would not be appropriate for predictive analysis. However, my object here is the estimation of the marginal effects from the conditional expectation function. For this purpose, the LPM is typically more appropriate (Angrist and Pischke 2008: pp. 94-106). In particular, while logit and probit models may have the benefit of bounding estimates of the dependent variable between 0 and 1, the resulting estimates may be biased by the incidental parameter problem (Neyman and Scott 1948). This may be particularly problematic when including many fixed effects as is necessary in this case.

Descriptive Statistics

Table 1a and 1b present descriptive statistics of investigators and grant cycles, respectively. In fitting with the very high levels of competition largely from established investigators to win an R01-equivalent grant, the sample is comprised of accomplished scientists. Investigators started the grant cycle at a mean of 18.7 years after completing their terminal degree with 56 prior publications and 2,234 career citations. A full 5% of the investigators achieve elite status during their careers, which I define as having ever won a Nobel Prize, Lasker Award, Howard Hughes Medical Institute Investigatorship, or membership in the Institute of Medicine or National Academy of Sciences. The majority of an investigator's research output is not directly supported by the focal R01-equivalent grant of interest. Among publications during the cycle period, 39.9% directly acknowledge the focal R01-equivalent NIH grant; 59.6% of publications which do not draw support from the focal NIH grant still had NIH support, whether from a non-R01-equivalent NIH grant to the investigator or a grant to one of their coauthors. Overall 45.2% of grant cycles were renewed; the renewal rate was lower for the first cycle of a grant (40.4%) than for subsequent cycles (51.5%).¹⁴ Grant cycles that were renewed had investigators that were younger (17.6 vs. 19.6 years), greater grant research productivity (8.1 vs. 7.0 publications and 663 vs. 422 citations among papers directly acknowledging the focal grant), and a greater percentage of the investigator's research activity during the cycle attributable to the focal grant (46% vs. 38.2% of all publications during the cycle had funding from the focal grant). Figure 1 provides the full distribution of the number of publications per grant cycle by renewal status.

Table 1c provides descriptive statistics for each of the four measures of risk taking and for novelty. While for most measures the mean level of risk taking was higher among grant cycles that were not renewed than those that were, the absolute differences are small. Figure 2 shows the full distribution of each of these measures by grant cycle renewal status. This provides initial descriptive evidence that the differences in renewal rate are not evenly distributed across different levels of risk taking for each of these measures. Table 1d presents the correlation for measures of risk taking and novelty. The relatively low correlation across

¹⁴ In contrast, the success rate of applicants for a new R01-equivalent grant is lower. In 2015, 16.1% of applications for a new R01-equivalent grant were funded (Lauer 2016).

these measures emphasizes how each is capturing a distinct approach to risk taking and how novelty contrasts with risk taking.

4. Results

Risk taking and grant renewal

Table 2 reports estimates for the impact of risk taking by each of my four measures on grant cycle renewal. Overall, I observe that high levels of risk are associated with lower renewal rates, and conversely, that low levels of risk are associated with higher renewal rates across each of these measures. Not only is the effect highly statistically significant, but the magnitude is substantial as well. For instance, when compared to grant cycles in the 10-90%ile for extreme tail outcomes, cycles in the bottom decile, reflecting low levels of risk taking, have a 5.1% higher renewal rate (an 11.0% increase) while those over the top decile have a 4.4% lower renewal rate (a 9.5% decrease) (Table 2A, column 2). Similar patterns are seen across the other measures of risk taking for the top and bottom deciles, including disruption (8.9% higher [19.6% increase] vs. 2.2% lower renewal rate [4.8% decrease]), pivoting (5.3% higher [11.6% increase] vs. 2.4% lower renewal rate [5.3% decrease]), and standing out (3.8% higher [7.7% increase] vs. 2.3% lower renewal rate [4.7% decrease]).¹⁵ These findings are robust to inclusion of controls for publication quantity and quality during a grant cycle, although in some cases this reduces the magnitude of the risk penalty. I do not find evidence that the risk-taking penalty decreases in subsequent grant cycles after a grant is initially renewed.

I also explore the full distribution of different levels of risk taking for its association with grant renewal (Figures 3, 4, 5, and 6). I estimate specifications in which there is a separate indicator variable for each ventile of the risk measure. Similar to Table 2, I find higher renewal rates for lower levels of risk taking and lower renewal rates for higher levels of risk taking. Across several of these measures, I find the striking pattern of a nearly monotonic decrease in renewal rates with increasing levels of risk taking.¹⁶

Novelty and risk taking

I next explore empirically the connection between risk taking and novelty. Figure 7 presents these results. I estimate specifications that are similar to Figures 2-5, but in addition to an indicator variable for the ventile of the measure of risk taking of interest I also include an indicator variable for each ventile of novelty. Two findings emerge from this analysis. First, in contradistinction to my measures of risk taking, I find novelty is rewarded. Additionally, the magnitude of the reward for novelty is similar to the magnitude

¹⁵ For extreme tail outcomes, pivoting, and standing out, percentiles are calculated relative to other grant cycles in the same evaluation year. For disruption, I emphasize work that is disruptive in an absolute sense and so use the percentile of the most disruptive paper, calculated relative to all publications in PubMed in the same year, for the grant cycle.

¹⁶ An exception to this is with disruption, where there is a non-linearity in the association between disruption and grant renewal: grant cycles with the very highest degree of disruption have higher renewal rates than those with elevated but less extreme levels of disruption.

of the risk-taking penalty. Second, inclusion of novelty does not significantly reduce the observed penalty for risk taking. This provides empirical evidence that novelty and measures of risk taking may have distinct dynamics, consistent with the idea that in some settings novelty is not a valid proxy for risk taking. Appendix Figure A2 repeats this analysis, but measures the effect of novelty on grant renewal without the inclusion of measures of risk taking. Similar to Figure 7, I find that novelty is associated with a higher probability of grant renewal in this analysis.

I also look at if the magnitude of the penalty for risk taking is different for novel research. Across all four measures of risk taking, grant cycles with above median novelty have a larger penalty for risk taking than those that are less novel (Table 3). At higher thresholds of novelty, the penalty for risk taking increases for pivoting and standing out, but is not statistically different from less novel research for the extreme tail outcomes and disruption measures (Appendix Table A1).

Who is penalized for risk taking?

I also explore heterogeneity in who is penalized for risk taking, with a focus on informing risk-taking dynamics as investigators progress through their career. I first focus on career age, measured as the time from terminal degree until grant cycle proposal. I then repeat the analysis in Figures 3-6, but interact indicator variables for the decile of each risk measure with an indicator variable for either being below or above the median career age. I similarly examine heterogeneity by gender, elite status as proxied by having won a prominent biomedical research award prior to grant cycle renewal, and grantsmanship as measured by the level of career NIH funding prior to the current grant cycle.

Figures 8 present these heterogeneity results for risk taking by extreme tail outcomes.¹⁷ Integrating across them, several trends emerge. First, risk taking is penalized across each subgroup examined. There are no subgroups explored for which risk taking wasn't penalized across each of the four measures. Second, early career investigators seem to face a larger penalty than their older colleagues for the same level of risk taking. Finally, some of these results are imprecisely estimated. This is especially the case for heterogeneity by status given the small number of elite investigators.

Robustness analysis and ancillary results

I also perform a number of robustness checks to assess the sensitivity of my estimates to alternative assumptions and subsamples. First, I extensively test alternative ways of measuring each of my approaches to risk taking and novelty (Appendix Table A2). In all cases, the results are robust to these alternative measures. This includes the difference between the maximum and median, median and minimum, winsorized maximum and minimum, and variance in vintage-adjusted citation percentiles for extreme tail outcomes (Appendix Figures A3 and A4); the mean rather than the maximum disruption index and a variation of the

¹⁷ Similar patterns are seen for heterogeneity measuring risk taking by disruption, pivoting and standing out.

index following Bornmann et al. (2020) for disruption (Appendix Figures A5 and A6); fraction new single MeSH rather than MeSH pairs and cosine dissimilarity score for single MeSH terms after term frequency-inverse document frequency weighting for pivoting (Appendix Figures A7 and A8); fraction unique single MeSH rather than MeSH pairs and cosine dissimilarity score for single MeSH terms after term frequency-inverse document frequency weighting for standing out (Appendix Figures A9 and A10), and mean MeSH and MeSH pair age for novelty (Appendix Figures A1 and A2).^{18,19} Emphasizing how pivoting and standing out are distinct from novelty, the findings with these measures are robust to exclusion of MeSH terms and MeSH term pairs which were first used in PubMed within the past decade.

The penalty for pivoting might be smaller for those grants for which the investigator proposed—and peer review approved and funded—to pivot in the initially submitted planned research studies. To examine this, I identify the fraction of MeSH terms shared between the grant proposal abstract and an investigator’s prior work and split the sample at the median for this measure. I find overall a similar magnitude penalty for pivoting across both groups (Appendix Table A3). In other words, grants which proposed a pivot and then investigators actually pivoted in the work carried out are still renewed at lower rates.

Author-ity has been shown to be highly accurate (Lerchenmueller and Sorenson 2016). It is likely the case, however, that its accuracy of publication disambiguation is higher for rare names within PubMed. To ensure my results are not driven by investigators with poorly disambiguated publications in Author-ity, I split the sample at the median name frequency based on frequency within the entire corpus of PubMed and find similar results across both groups (Appendix Table A4).

One reason investigators may choose not to renew their grant is that their primary research focus lies outside of the scope of what the NIH typically supports. For instance, some organic chemists, electrical engineers, or sociologist investigators may have an isolated project at the interface of biomedicine within the context of a broader research agenda elsewhere. To ensure my results are not driven by investigators with only a weak attachment to NIH funding, I calculate the fraction of career publications with any NIH funding and split the sample at the median of this measure (Appendix Table A5). I find a similar penalty for risk taking across both those with above and below the median career attachment to NIH funding.

One might be concerned, especially for articles published well before the digital era, that not all articles that in fact received NIH funding are recorded as having acknowledged it. Care must be taken when examining heterogeneity over time, as the degree of competition for R01-equivalent grants has been increasing and NIH policy has evolved over time. To reassure that my results are not driven by this potential

¹⁸ Inverse document frequency was performed relative to all original research publications in PubMed in the year the grant cycle came up for renewal.

¹⁹ Since for standing out I am interested in identifying investigators whose research substantially departs from other investigators, as a robustness check, I repeat the analysis limiting it only to investigators with at least 10 other funded researchers with abstracts available in the same IC-study section-year with a larger risk penalty observed.

measurement error, I split the sample at the median grant cycle evaluation year and find similar results across both groups (Appendix Table A6).

Finally, I also use a bounding technique proposed by Oster (2019) to gauge the sensitivity of my results to a failure of the unconfoundedness assumption. The intuition behind this approach is that the stability of the coefficient for risk taking when varying the set of control variables included in the model, scaled by the 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 outcomes 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 δ , which can be interpreted as the degree of selection on unobservables relative to observables necessary to reduce the magnitude of the effect of the regressor of interest to zero. Appendix Table A7 reports the results of this exercise. For extreme tail outcomes, disruption, and pivoting, δ is far above one, the threshold recommended by Oster to suggest robustness to the influence of unobservable covariates, with standing out having a δ of -0.70 and -0.59.²⁰

5. Discussion

Using the universe of awarded NIH R01-equivalent grants and across a range of different measures of risk taking, I find that investigators with high levels of risk taking are renewed at lower rates than those with low levels of risk taking. In this respect, these findings provide systematic evidence that peer review at the NIH punishes risk taking. I also empirically demonstrate that novelty has distinct dynamics from my measures of risk taking, emphasizing that in some settings it is not a valid proxy for risk taking. In addition to informing peer review and innovation project selection processes more generally, this is a population of substantial policy interest given the prominence of the NIH and the R01 grant mechanism in funding biomedical research in the United States.

A strength of my approach is its focus on measuring the association between risk taking and grant renewal across the full distribution of observed levels of risk taking. This emphasis on the tail-values of risk taking reflects that not all scientific innovation impact scientific progress equally, and it may be those with the largest degree of risk are most likely to be associated with radical innovation (Rzhetsky et al. 2015, Veugelers and Wang 2019, Wang et al. 2017). It is also striking that peer review punished risk taking in light of the mixed evidence on the ability of peer review scores to predict scientific productivity and impact (Cole et al. 1981, Danthi et al. 2014, Li and Agha 2015, Pier et al. 2018, Rothwell and Martyn 2000).²¹

²⁰ Some values for δ are negative. This reflects that the correlation between unobservables and risk taking would need to be in the opposite direction as the correlation between observables and risk taking to explain away the result.

²¹ This parallels the dynamics seen for innovation in other settings, such as entrepreneurship. For instance, even among VC-backed startups, a handful of companies that are outliers in their success account for the majority of VC profits and impact on the larger economy. Likewise, while investors may be able to identify those companies with little growth potential, they have only a very limited ability to identify among VC-funded companies which ones are likely to have outlier success (Kerr et al. 2014).

The validity of my results relies upon the maintained assumption there is no significant omitted variable bias: that is, that there are no covariates associated with both risk taking and grant renewal which are not included in my empirical strategy. I am reassured this is not a large concern given the rich set of controls I employ as well as how pervasive and robust my findings are that risk taking is penalized. Additionally, there are also several reasons to think the estimates may reflect a lower bound of the possible risk penalty. First, several of my approaches to measuring risk rely on the presence of publications. To the extent high risk taking is more likely to result in the absence of publications and that the absence of publications leads to a low renewal rate, measuring risk taking conditional on having publications may underestimate the total penalty for risk taking. Second, as I am unable to observe unfunded grant applications, entry into my sample is conditional upon having first applied for and received an NIH R01-equivalent grant. Any risk-taking punishment in winning an initial NIH grant would lead to my sample having lower overall levels or less extreme distributions of risk taking than the population of initial NIH R01-equivalent applicants. This suggests the magnitude of risk-taking penalization in this study is lower than I might expect if I could repeat the study using initial applications for NIH R01-equivalent grants. Similarly, investigators who prioritize high levels of risk taking in their research, aware of the NIH's reputation for penalizing risk, may not seek NIH funding given the time and opportunity costs involved in applying (Langer 2012). This would represent another indirect way grant peer review penalizes risk taking not captured in my study. It is also possible that researchers in training—many of whom work under the mentorship of NIH-funded investigators—may imprint with a less risky approach to research after observing their mentors avoid proposing risky projects for NIH funding (Azoulay et al. 2021, Higgins 2005).

I argue each measure of risk taking captures of a different aspect of what it means to perform risky research, each with its accompanying strengths and weaknesses. For instance, an individual scientist exploring a new area relative to their prior research may reflect promising recombination of ideas, or alternatively it may reflect devaluing of the scientist's prior experience with a decreased likelihood of success and lower ultimate impact (Arts and Fleming 2018). Similarly, scientists who pursue research few other scientists are interested in may reflect potential breakthrough innovation, or it may reflect simply a less promising research question. The consistency of my results across these different measures is reassuring against the particular weaknesses of each individual measure that what I am capturing, in fact, reflects scientific risk taking. As it is generally very difficult to publish failed experiments and null results, that my findings are exclusively based on risk taking measured from successfully published work is reassuring that the results are not driven by grant non-renewal after a risky project resulted in a failed experiment or null outcome.

There are many potential mechanisms through which peer review may act to penalize risk. This study does not differentiate between the relative contributions of these various mechanisms towards the observed risk penalty. Similarly, while I observe which grants are renewed, I do not observe if they were

not renewed because an investigator chose not to apply or applied but was not selected. I expect exploring these issues to be fruitful areas for future research.

Both high-risk, high-reward and low-risk, incremental projects are valuable and play a role in advancing the scientific frontier. Yet there are reasons to think the NIH playing a larger role supporting high-risk research could be valuable. In particular, one would expect that established, mature lines of scientific inquiry should contain more incremental research—and less risk taking—than novel, cutting-edge lines of scientific inquiry which would be best served by greater exploration and risk taking. Instead, I find the opposite: the penalty for risk taking is larger for research that is more novel. This argues that the current degree of penalization for risk taking at the NIH results in a welfare loss. In addition, if the NIH penalizes risk taking, this may have a disproportionate impact on research for which other parties have limited incentives to provide. Public investments in early-stage basic research are often justified with the positive knowledge spillovers and difficulty for private industry to appropriate its value (Aghion et al. 2008). Indeed, the government is precisely the agent best positioned to bear risks that are too large for other individuals or organizations to bear. Corporations have also had decreasing investments in science in the recent decades (Arora et al 2017, Fleming et al. 2019). This, along with the NIH being a quantitatively important funder of research due to its size and scope, suggests hesitancy by the NIH to fund risky early-stage basic science research may not easily be compensated for by other funding sources and may lead to a slower overall pace of technical advancement (Foster et al. 2015). Similarly, the NIH may play an important role in sponsoring clinical studies for which there is limited ability for industry to appropriate value, resulting in a potential under-provision of risk taking in this type of research (Greenblatt et al. 2023). While private philanthropies provide an alternative source of research funding, their priorities and areas of support may diverge from those with the most compelling public benefit (Murray 2012).²²

The challenge of supporting high-risk research has received increasing attention in recent years, including from the NIH. The NIH’s efforts have often centered on creating distinct funding mechanisms with explicit mandates around high-risk, high-reward research. This includes the Pioneer Awards, Director’s Transformative R01 Awards, and New Innovator Awards (NIH 2019). In 2021, the NIH funded 106 such awards with an expected accompanying outlay of \$329 million over five years (NIH 2021).²³ A number of private philanthropic science funders have dedicated opportunities to support high-risk research, including

²² Researchers who do not receive NIH funding may be able to substitute their research with other fund sources. For instance, Chuban and Hackett (1990) find that many investigators are able to pursue their proposed research even if they do not receive the focal grant. Similarly, Jacob and Lefgren (2011) find the marginal impact of receiving an NIH grant to be only about one additional publication over five years, in part because researchers can substitute to other funding sources. Yet NIH punishing risk taking may have an indirect effect on what proposals investigators take to these sources. It remains an open question the extent to which these other sources themselves punish scientific risk taking.

²³ This represents 1.4% by number of awards and 1.8% by funding amount of all R01-equivalent awards that year at the NIH (Lauer 2022).

the Howard Hughes Medical Institute, Open Philanthropy Project, Fast Grants, Chan Zuckerberg Initiative, and Wellcome Trust.

While the focus here has been on how peer review shapes selection of projects by public funders, my work has implications for the selection of high-risk innovation projects in other settings as well. First, rather than being considered as an afterthought or as simply words of admonition to project evaluators, how to support the desired level of risk should be built into the design of supporting processes and institutions. While dedicated funding is an important step forward, it still must contend with bias against novelty or an over-reliance on criteria that are easy to measure but may be poorly correlated with promising risk taking. For instance, as high-risk projects may have greater variation in evaluator assessments of quality, moving from a consensus-based to a champion-based approach or emphasizing the degree of enthusiastic support rather than mean score may facilitate risk taking (Azoulay and Li 2022, Franzoni et al. 2022). Processes which encourage independent assessment of projects by the entire population of evaluators may help better identify promising risky proposals than those in which only a subset of the panel evaluates each application in detail. Similarly, who comprises the evaluation panel may affect the evaluation. For instance, expert evaluators may have more pronounced bias against novelty in their own field (Wang et al. 2017). Having greater diversity of expertise and traditions, as well as explicitly considering an evaluator’s own prior risk taking behavior, when consider whom to appoint to select innovation proposals may also reduce risk penalization (Nicholson and Ioannidis 2012). Finally, a greater emphasis on selection and outcomes evaluation at the level of the portfolio of projects rather than the individual project level may facilitate greater acceptance of strategic risk taking.

Second, a fundamental challenge of risky innovation is ex post, a failed but highly promising risky research project and a low-quality idea can look much the same, with no publication, patent, or product to show for the effort. While innovators can be incentivized by status and financial awards for successful risky innovation, they may be reluctant to do so when failed risky projects might lead to job loss or lab closure. While it is important to maintain high standards and ultimate accountability (Manso 2011), this may mean adjusting the criteria upon which innovators are evaluated to focus more on innovation inputs and the level of risk rather than simply the outputs of innovative efforts.

Finally, I encourage greater efforts to build evaluation into the design of programs to support risk taking. Such efforts to turn the scientific method upon the design of institutions and processes to support risk taking is likely to allow funders to more effectively develop the portfolio of projects most likely to meet their goals. I believe such efforts are a risk worth taking.

References

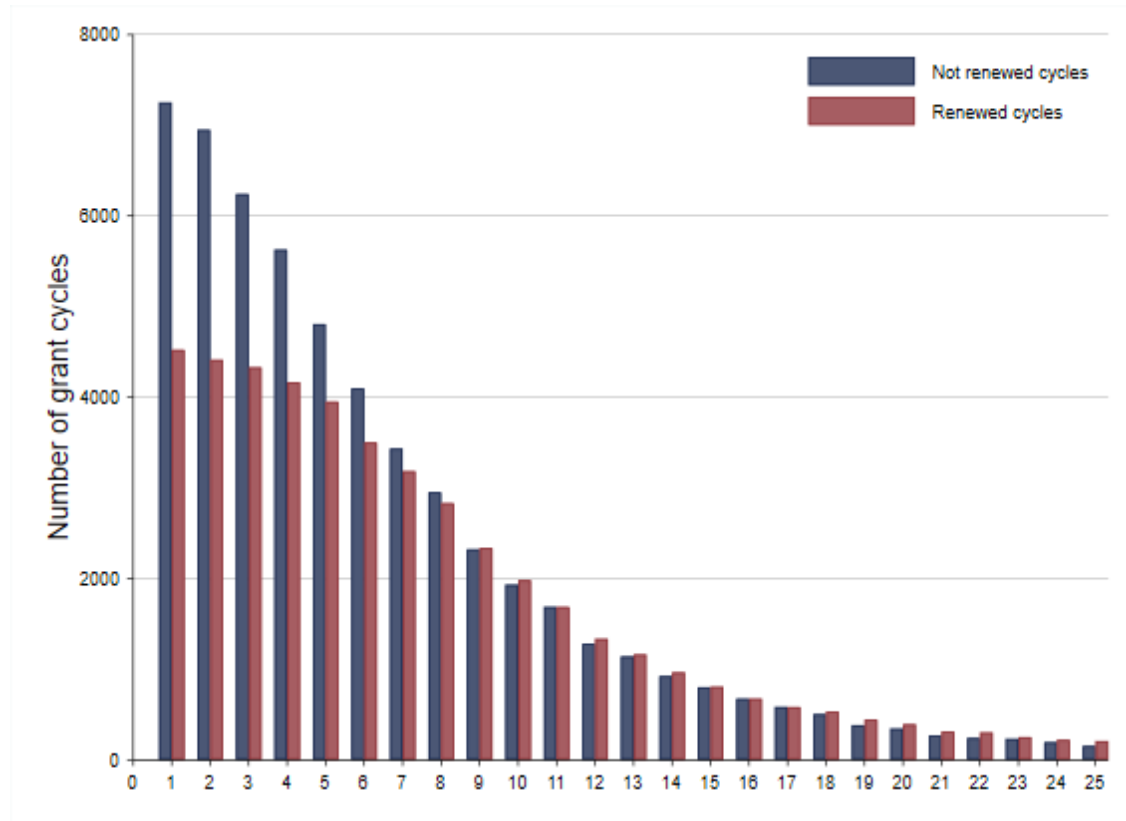
- Aghion, Philippe, Mathias Dewatripont, and Jeremy C. Stein. 2008. "Academic freedom, private-sector focus, and the process of innovation." *RAND Journal of Economics*, 39(3):617-635.
- Alberts, Bruce, Marc W. Kirschner, Shirley Tilghman, and Harold Varmus. 2014. "Rescuing US biomedical research from its systemic flaws." *Proceedings of the National Academy of Sciences*, 111(16):5773-5777.
- Althaus, Catherine E. 2005. "A disciplinary perspective on the epistemological status of risk." *Risk Analysis*, 25(3):567-588.
- Angrist, Joshua D., and Jorn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.
- Arora, Ashish, Sharon Belenzon, and Andrea Pataconi. 2017. "The decline of science in corporate R&D." *Strategic Management Journal*, 39(1):3-32.
- Arts, Sam, and Lee Fleming. 2018. "Paradise of Novelty—Or Loss of Human Capital? Exploring New Fields and Inventive Output" *Organization Science*, 29(6):1074-1092.
- Ayoubi, Charles, Michele Pezzoni, and Fabiana Visentin. 2021. "Does it Pay to do Novel Science? The Selectivity Patterns in Science Funding." *Science and Public Policy*, 48(5):635-648.
- 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.
- Azoulay, Pierre, Joshua S. Graff Zivin, and Gustavo Manso. 2011. "Incentives and creativity: evidence from the academic life sciences." *RAND Journal of Economics*, 42(3):527-554.
- Azoulay, Pierre, Joshua S. Graff Zivin, Danielle Li, and Bhaven Sampat. 2019. "Public R&D and Private-Sector Patenting: Evidence from NIH Funding Rules." *The Review of Economic Studies*, 86(1):117-152.
- Azoulay, Pierre, Wesley H. Greenblatt, and Misty L. Heggeness. 2021. "Long-term effects from early exposure to research: Evidence from the NIH 'Yellow Berets.'" *Research Policy*, 50(9):104332.
- Azoulay, Pierre and Danielle Li. 2022. "Scientific Grant Funding." In Goolsbee, Austan and Benjamin Jones (Eds). *Innovation and Public Policy*. Chicago, IL:University of Chicago Press.
- Bachrach, C.A., and Thelma Charen. 1978. "Selection of Medline contents, the development of its thesaurus, and the indexing process." *Medical Informatics*, 3(3):237-254.
- Banal-Estanol, Albert, Ines Macho-Stadler, and David Perez-Castrillo. 2019. "Evaluation in research funding agencies: Are structurally diverse teams biased against?" *Research Policy*, 48(7):1823-1840.
- Bornmann, Lutz, Sitaram Devarakonda, Alexander Tekles, and George Chacko. 2020. "Are disruption index indicators convergently valid? The comparison of several indicator variants with assessments by peers." *Quantitative Science Studies*, 1(3):1242-1259.
- Boudraeu, Kevin J., Eva C. Guinan, Karim R. Lakhani, and Christoph Riedl. 2016. "Looking Across and Looking Beyond the Knowledge Frontier: Intellectual Distance, Novelty, and Resource Allocation in Science." *Management Science*, 62(10):2765-2783.
- Bromham, Lindell, Russell Dinnage, and Xia Hua. 2016. "Interdisciplinary research has consistently lower funding success." *Nature*, 534(7609):684-687.
- Carson, Richard T., Joshua Graff Zivin, Jordan J. Louviere, Sally Sadoff, and Jeffrey G. Shrader. 2022. "The Risk of Caution: Evidence from an Experiment." *Management Science*, 68(12):9042-9060.
- Chubin, Daryl E., and Edward J. Hackett. 1990. *Peerless Science: Peer Review and U.S. Science Policy*. Albany, NY: SUNY Press.
- Cole, Stephen, Jonathan R. Cole, and Gary A. Simon. 1981. "Chance and Consensus in Peer Review." *Science*, 214(4523):881-886.
- Coletti, Margaret H. and Howard L. Bleich. 2001. "Medical Subject Headings Used to Search the Biomedical Literature." *Journal of the American Medical Informatics Association*, 8(4):317-323.
- Collison, Patrick, Tyler Cowen, and Patrick Hsu. 2021. "What We Learned Doing Fast Grants." *Future*.
- Danthi, Narasimhan, Colin O. Wu, Peibei Shi, and Michael Lauer. 2014. "Percentile ranking and citation impact of a large cohort of National Heart, Lung, and Blood Institute-funded cardiovascular R01 grants." *Circulation Research*, 114(4):600-606.

- Dasgupta, Partha, and Paul A. David. 1994. "Toward a new economics of science." *Research Policy*, 23:487-521.
- Fang, Ferric C., and Arturo Casadevall. 2015. "Competitive science: Is competition ruining science?" *Infection and Immunity*, 83(4):1229-1233.
- Fields, Stanley. 2014. "Would Fred Sanger get funded today?" *Genetics*, 197(2):435-439.
- Fleming, Lee. 2001. "Recombinant Uncertainty in Technological Search." *Management Science*, 47(1):117-132.
- Fleming, Lee, Hillary Greene, Guan-Cheng Li, Matt Marx, and Dennis Yao. 2019. "Government-funded research increasingly fuels innovation." *Science*, 364(6446):1139-1141.
- Foster, Jacob G., Andrey Rzhetsky, and James A. Evans. 2015. "Tradition and Innovation in Scientists' Research Strategies." *American Sociological Review*, 80(5):875-908.
- Fox, Craig R., and Amos Tversky. 1995. "Ambiguity Aversion and Comparative Ignorance." *Quarterly Journal of Economics*, 110(3):585-603.
- Franzoni, Chiara, and Paula Stephan. 2023. "Uncertainty and risk-taking in science: Meaning, measurement and management in peer review of research proposals." *Research Policy*, 52(3):104706.
- Franzoni, Chiara, Paula Stephan, and Reinhilde Veugelers. 2022. "Funding Risky Research." *Entrepreneurship and Innovation Policy*, 1:103-133.
- Funk, Mark E., and Carolyn Anne Reid. "Indexing consistency in MEDLINE." *Bulletin of the Medical Library Association*, 71(2):176-183.
- Funk, Russell J., and Jason Owen-Smith. 2017. "A Dynamic Network Measure of Technological Change." *Management Science*, 63(3):791-817.
- Gerin, William, Christine K. Kinkade, and Niki L. Page. 2018. *Writing the NIH Grant Proposal: A Step-by-Step Guide*. Thousand Oaks, CA: Sage Publications.
- Graff Zivin, Joshua, and Elizabeth Lyons. 2021. "The Effects of Prize Structures on Innovative Performance." *AEA Papers and Proceedings*, 111:577-581.
- Greenblatt, Wesley, Charu Gupta, and Jennifer Kao. 2023. "Drug Repurposing During the COVID-19 Pandemic: Lessons for Expediting Drug Development and Access." *Health Affairs*, 42(3):424-432.
- Higgins, Monica C. 2005. *Career Imprints: Creating Leaders Across an Industry*. San Francisco: Jossey-Bass/Wiley.
- Hill, Ryan, Yian Yin, Carolyn Stein, Dashun Wang, and Benjamin F. Jones. 2021. "Adaptability and the Pivot Penalty in Science." Working paper.
- Jacob, Brian A. and Lars Lefgren. 2011. "The Impact of research grant funding on scientific productivity." *Journal of Public Economics*, 95(9-10):1168-1177.
- Kent, David. 2018. "Scientists can't take risks until they are in their 40s." *University Affairs*.
- Kerr, William R., Ramana Nanda, and Matthew Rhodes-Kropf. 2014. "Entrepreneurship as Experimentation." *Journal of Economic Perspectives*, 28(3):25-48.
- Krieger, Joshua, Danielle Li, and Dimitris Papanikolaou. 2022. "Missing Novelty in Drug Development." *The Review of Financial Studies*, 35(2):636-679.
- Kolata, Gina. 2009. "Grant System Leads Cancer Researchers to Play it Safe." *New York Times*.
- Kornberg, Roger D. 2007. "2006 Nobel Laureates." U.S. Senate Committee on Commerce, Science, and Transpiration.
- Krieger, Joshua, and Ramana Nanda. 2022. "Are Transformational Ideas Harder to Fund? Resource Allocation to R&D Projects at a Global Pharmaceutical Firm." Harvard Business School Working Paper No. 23-014.
- Kuhn, Thomas S. 1962. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.
- Lane, Jacqueline N., Zoe Szajnfarter, Jason Crusan, Michael Menietti, and Karim R. Lakhani. 2022a. "Are Experts Blinded by Feasibility? Experimental Evidence from a NASA Robotics Challenge." Harvard Business School Working Paper No. 22-071.
- Lane, Jacqueline N., Misha Teplitskiy, Gary Gray, Hardeep Ranu, Michael Menietti, Eva C. Guinan, and Karim R. Lakhani. 2022b. "Conservatism Gets Funded? A Field Experiment on the Role of Negative Information in Novel Project Evaluation." *Management Science*, 68(6):4478-4495.
- Langer, James S. 2012. "Enabling Scientific Innovation." *Science*, 338(6104):171.
- Lauer, Mike. 2022. "FY 2021 by the Numbers: Extramural Grant Investments in Research." National Institutes of Health Office of Extramural Research.

- Lauer, Mike. 2016. "Are Attempts at Renewal Successful?." National Institutes of Health Office of Extramural Research.
- Lee, You-Na, John P. Walsh, and Jian Wang. 2015. "Creativity in scientific teams: Unpacking novelty and impact." *Research Policy*, 44(3):684-697.
- Lercehmueller, Marc J. and Olav Sorenson. 2016. "Author Disambiguation in PubMed: Evidence on the Precision and Recall of Author-ity among NIH-Funded Scientists." *PLoS One*, 11(7):e0158731.
- Lerner, Josh and Julie Wulf. 2007. "Innovation and Incentives: Evidence from Corporate R&D." *Review of Economics and Statistics*, 89(4):634-644.
- Li, Danielle and Leila Agha. 2015. "Big names or big ideas: Do peer-review panels select the best science proposals?" *Science*, 348(6233):434-438.
- Lorsch, Jon R. 2015. "Maximizing the return on taxpayers' investments in fundamental biomedical research." *Molecular Biology of the Cell*, 26(9):1578-1582.
- Manso, Gustavo. 2011. "Motivating Innovation." *Journal of Finance*, 66(5):1823-1860.
- Mishra, Shubhanshu and Vetle I. Torvik. 2016. "Quantifying Conceptual Novelty and the Biomedical Literature." *Dlib Mag.*, 22(9-10).
- Moses 3rd, Hamilton, David H. M. Matheson, Sarah Cairns-Smith, Benjamin P. George, Chase Palisch, and E. Ray Dorsey. 2015. "The anatomy of medical research: US and international comparisons." *Journal of the American Medical Association*, 313(2):174-89.
- Mueller, Jennifer S., Shimul Melwani, and Jack A. Goncalo. 2012. "The Bias Against Creativity: Why People Desire but Reject Creative Ideas." *Psychological Science*, 23(1):13-17.
- Murray, Fiona. 2012. "Evaluating the Role of Science Philanthropy in American Universities." *Innovation Policy and the Economy*, 13:23-59.
- National Institutes of Health. 2021. "NIH supports 106 grants featuring high-risk, high-reward research." <https://www.nih.gov/news-events/news-releases/nih-supports-106-grants-featuring-high-risk-high-reward-research>. Press release.
- National Institutes of Health. 2019. "Report of the ACD Working Group on High-Risk, High-Reward Research." Advisory Committee to the Director Working Group on High-Risk High-Reward Programs, National Institutes of Health.
- Nielsen, Michel (michael_nielsen). 2022. "(A brief afternoon rant.)" Tweet.
- Neyman, Jerzy, and Elizabeth L. Scott. 1948. "Consistent Estimates Based on Partially Consistent Observations." *Econometrica* 16(1):1-32.
- Nicholson, Joshua M. and John P. A. Ioannidis. 2012. "Conform and be funded." *Nature*, 492(7427):34-36.
- Packalen, Mikko, and Jay Bhattacharya. 2020. "NIH funding and the pursuit of edge science." *Proceedings of the National Academy of Sciences*, 117(22):12011-12016.
- Petsko, Gregory A. 2012. "Goodbye, Columbus." *Genome Biology*, 13(5):155.
- Pier, Elizabeth L. Markus Brauer, Amarette Filut, Anna Kaatz, Joshua Raclaw, Mitchell J. Nathan, Cecilia E. Ford, and Molly Carnes. 2018. "Low agreement among reviewers evaluating the same NIH grant applications." *Proceedings of the National Academy of Sciences*, 115(2):2952-2957.
- Rosenberg, Nathan. 1995. "Uncertainty and Technological Change." *Federal Reserve Bank of Boston Conference Proceedings*, 40(Jun):91-125.
- Rothwell, Peter M., and Christopher N. Martyn. 2000. "Reproducibility of peer review in clinical neuroscience." *Brain*, 123:1964-1969.
- Ruben, Adam. 2017. "Another tenure-track scientist bites the dust." Experimental Error, *Science*, <https://www.science.org/content/article/another-tenure-track-scientist-bites-dust>.
- Rzhetsky, Andrey, Jacob G. Foster, Ian T. Foster, and James A. Evans. 2015. "Choosing experiments to accelerate collective discovery." *Proceedings of the National Academy of Sciences*, 112(47):14569-74.
- Schilling, Melissa A. and Elad Green. 2011. "Recombinant search and breakthrough idea generation: An analysis of high impact papers in the social sciences." *Research Policy*, 40(10):1321-1331.
- Shibayama, Sotaro, Deyun Yin, and Kuniko Matsumoto. 2021. "Measuring novelty in science with word embedding." *PLoS ONE*, 16(7):e0254034.
- Shibayama, Sotaro and Jian Wang. 2020. "Measuring originality in science." *Scientometrics*, 122:409-427.
- Sutter, Paul M. 2022. "Risk Aversion Is Ruining Science." *Undark*.
- Stephan, Paula E. 2012. *How Economics Shapes Sciences*. Cambridge, MA: Harvard University Press.

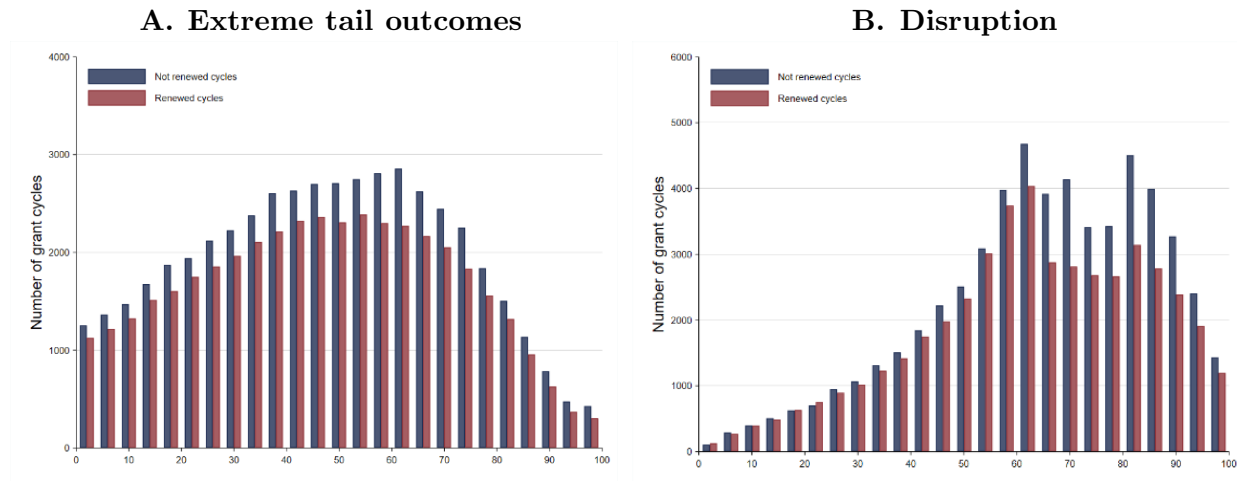
- Teplitskiy, Misha, Hao Peng, Andrea Blasco, and Karim R. Lakhani. 2022. "Is novel research worth doing? Evidence from peer review at 49 journals." *Proceedings of the National Academy of Sciences*, 119(47):e2118046119.
- Tian, Xuan and Tracy Yue Wang. 2014. "Tolerance for Failure and Corporate Innovation." *Review of Financial Studies*, 27(1):211-255.
- 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. and Neil R. Smalheiser. 2021. Author-ity 2018 - PubMed author name disambiguated dataset. University of Illinois at Urbana-Champaign.
- 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.
- Uzzi, Brian, Satyam Mukherjee Michael Stringer, and Ben Jones. 2013. "Atypical Combinations and Scientific Impact." *Science*, 342(6157):468-472.
- Verhoeven, Dennis, Jurrien Bakker, and Reinhilde Veugelers. 2016. "Measuring technological novelty with patent-based indicators." *Research Policy*, 45(3):707-723.
- Veugelers, Reinhilde, Jian Wang, and Paula Stephan. 2022. "Do Funding Agencies Select and Enable Risky Research: Evidence from ERC Using Novelty as a Proxy of Risk Taking." National Bureau of Economic Research Working Paper #30320.
- Veugelers, Reinhilde and Jian Wang. 2019. "Scientific novelty and technological impact." *Research Policy*, 48:1362-1372.
- Wang, Jian, Reinhilde Veugelers, and Paula Stephan. 2017. "Bias against novelty in science: A cautionary tale for users of bibliometric indicators." *Research Policy*, 46:1416-1436.
- Woolston, Chris. 2014. "Sanger's legacy stirs up digital chatter." *Nature*, 511(9).
- Wu, Lingfei, Dashun Wang, and James A. Evans. 2019. "Large teams develop and small teams disrupt science and technology." *Nature*, 566(7744):378-382.

Figure 1. Distribution of grant cycle publications by renewal status



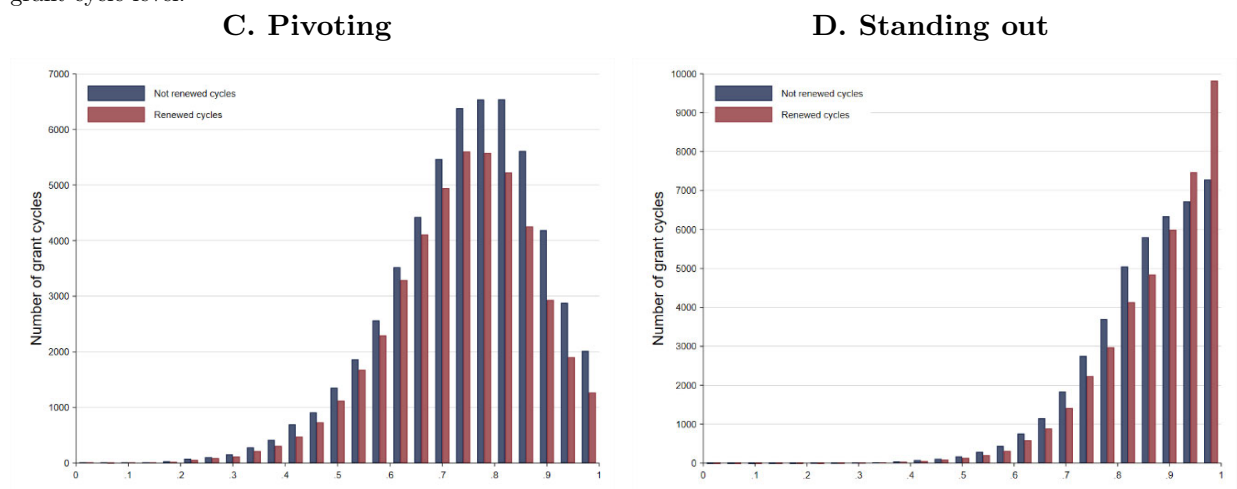
Note: Histogram of the number of publications acknowledging the grant within the grant cycle. Unit of analysis is at the investigator-grant-cycle level. Two thousand nine hundred fifty-two (2.9% of the sample) outliers with over 25 grant cycle publications are not shown.

Figure 2. Distribution of risk-taking measures by renewal status



Note: Histogram of range between the maximum and minimum vintage-adjusted citation percentile for grant cycle publications. Unit of analysis is at the investigator-grant-cycle level.

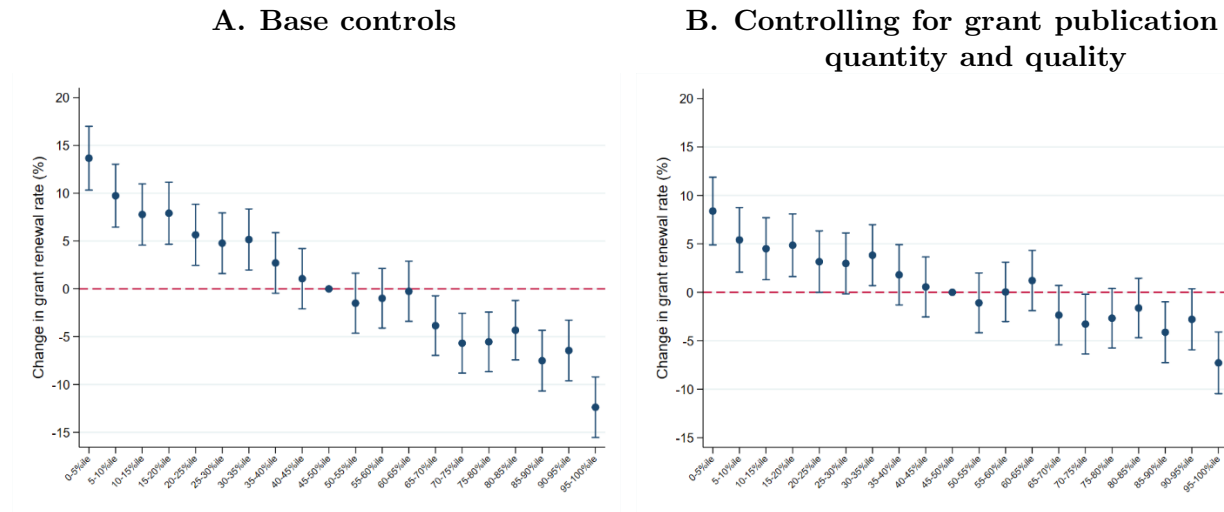
Note: Histogram of the maximum disruption index percentile among grant cycle publications. Unit of analysis is at the investigator-grant-cycle level.



Note: Histogram of the fraction of MeSH term pairs for grant cycle publications that were not used in the investigator's publications in the 5 years preceding the grant. Unit of analysis is at the investigator-grant-cycle level.

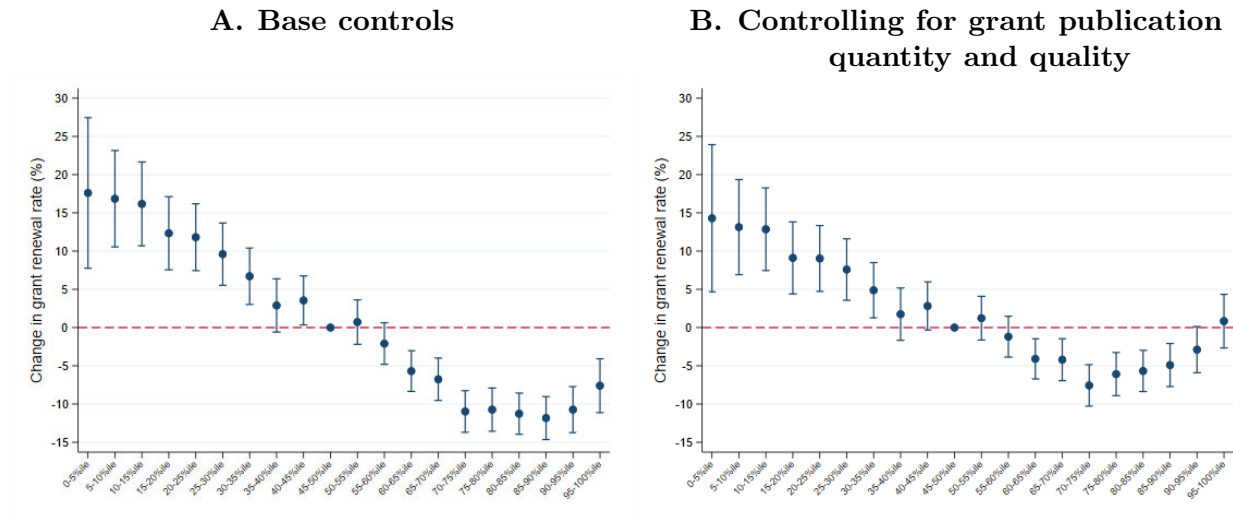
Note: Histogram of the fraction of MeSH term pairs for grant cycle publications that were not used by funded grant proposal abstracts in the same NIH IC-study section-year. Unit of analysis is at the investigator-grant-cycle level.

Figure 3. Effect of risk taking on grant renewal: Extreme tail outcomes



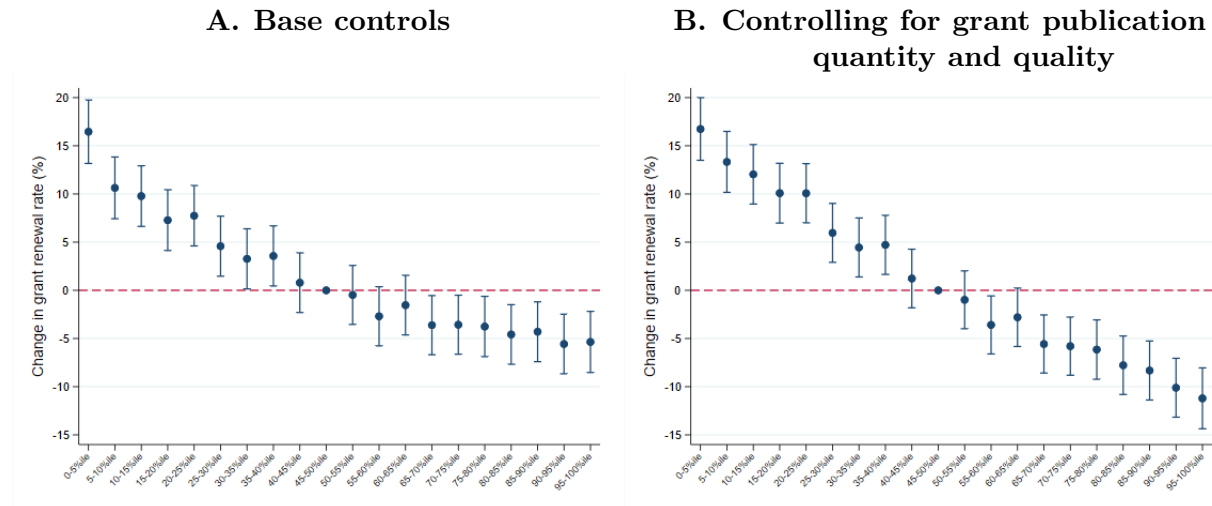
Note: The plots above present the coefficient estimates (and associated 95% confidence intervals) from a linear probability model where the dependent variable is an indicator variable for if the grant cycle was renewed. The unit of analysis is the investigator-grant-cycle. Panel A was modeled on Table 2, column 2, while panel B was modeled on Table 2, column 3. Each regression includes a set of 19 indicator variables for each ventile of the independent variable of interest, the difference between the maximum and minimum vintage-adjusted citation percentile among grant cycle publications. The omitted category is the 45-50%ile. Robust standard errors were used, clustered at the investigator and grant.

Figure 4. Effect of risk taking on grant renewal: Disruption



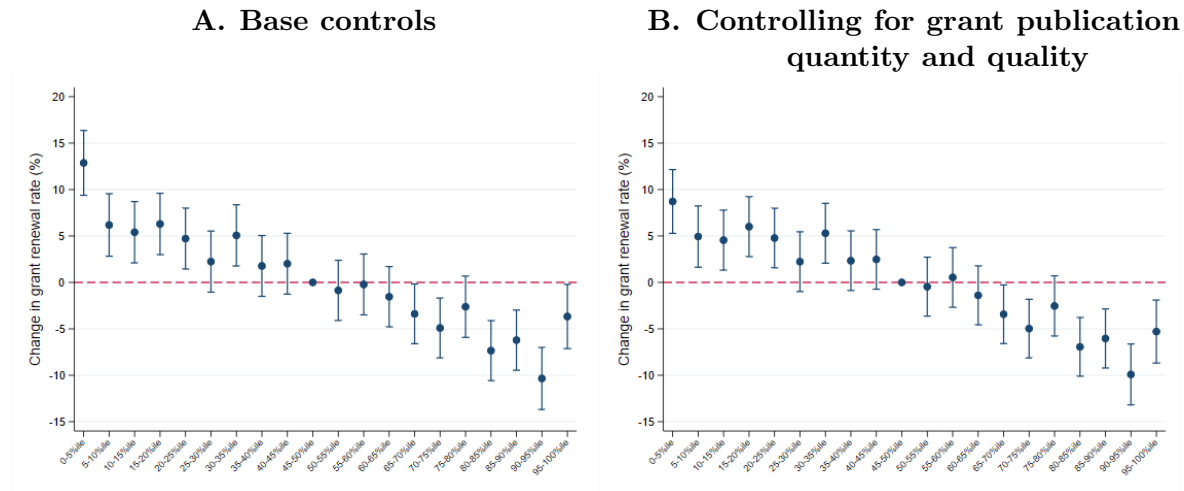
Note: The plots above present the coefficient estimates (and associated 95% confidence intervals) from a linear probability model where the dependent variable is an indicator variable for if the grant cycle was renewed. The unit of analysis is the investigator-grant-cycle. Panel A was modeled on Table 2, column 2, while panel B was modeled on Table 2, column 3. Each regression includes a set of 19 indicator variables for each ventile of the independent variable of interest, the maximum of the disruption index percentile for grant cycle publications. The omitted category is the 45-50%ile. Robust standard errors were used, clustered at the investigator and grant.

Figure 5. Effect of risk taking on grant renewal: Pivoting



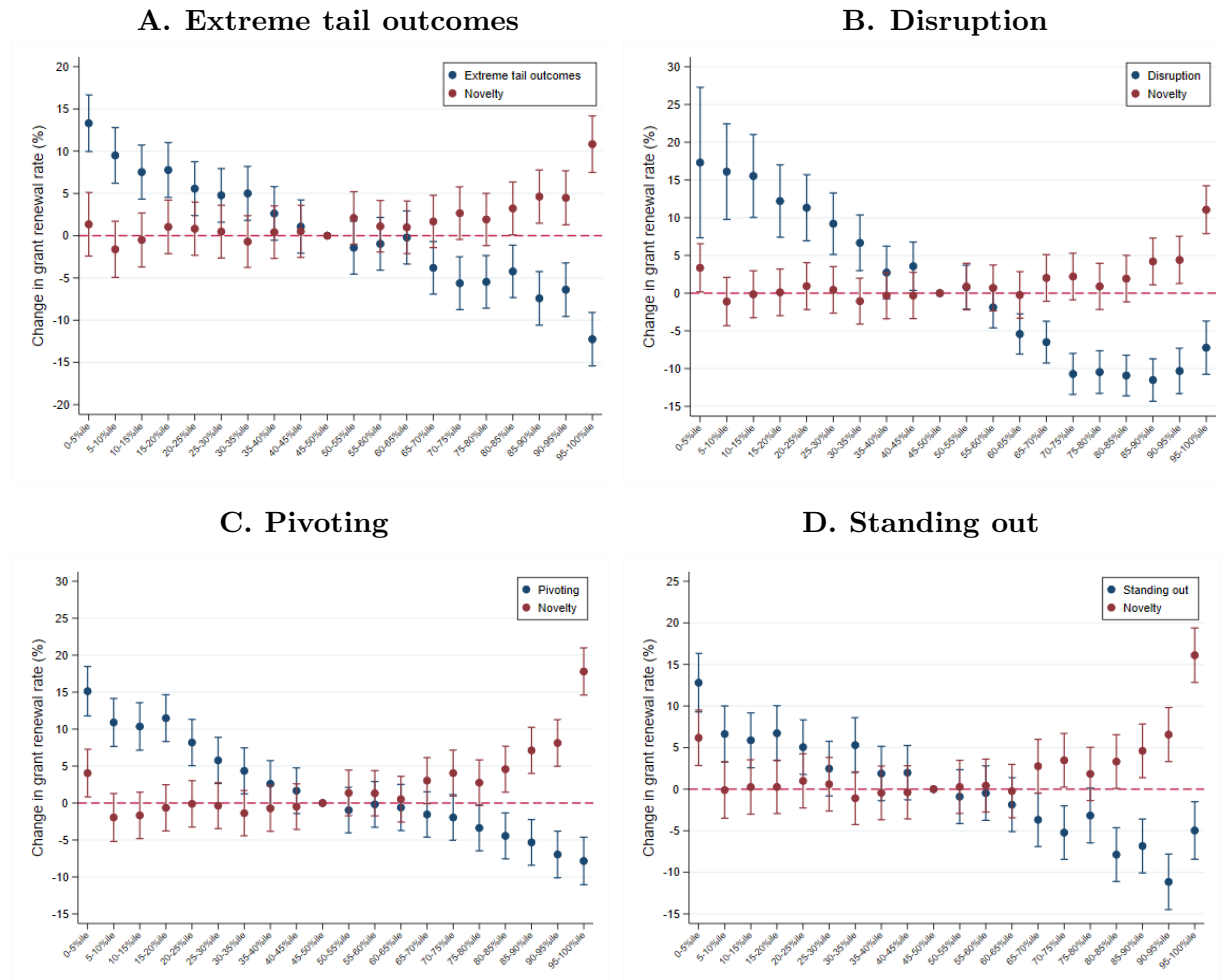
Note: The plots above present the coefficient estimates (and associated 95% confidence intervals) from a linear probability model where the dependent variable is an indicator variable for if the grant cycle was renewed. The unit of analysis is the investigator-grant-cycle. Panel A was modeled on Table 2, column 2, while panel B was modeled on Table 2, column 3. Each regression includes a set of 19 indicator variables for each ventile of the independent variable of interest, the fraction of MeSH term pairs from grant cycle publications that were not used by the investigator in the 5 years preceding the grant cycle. The omitted category is the 45-50%ile. Robust standard errors were used, clustered at the investigator and grant.

Figure 6. Effect of risk taking on grant renewal: Standing out



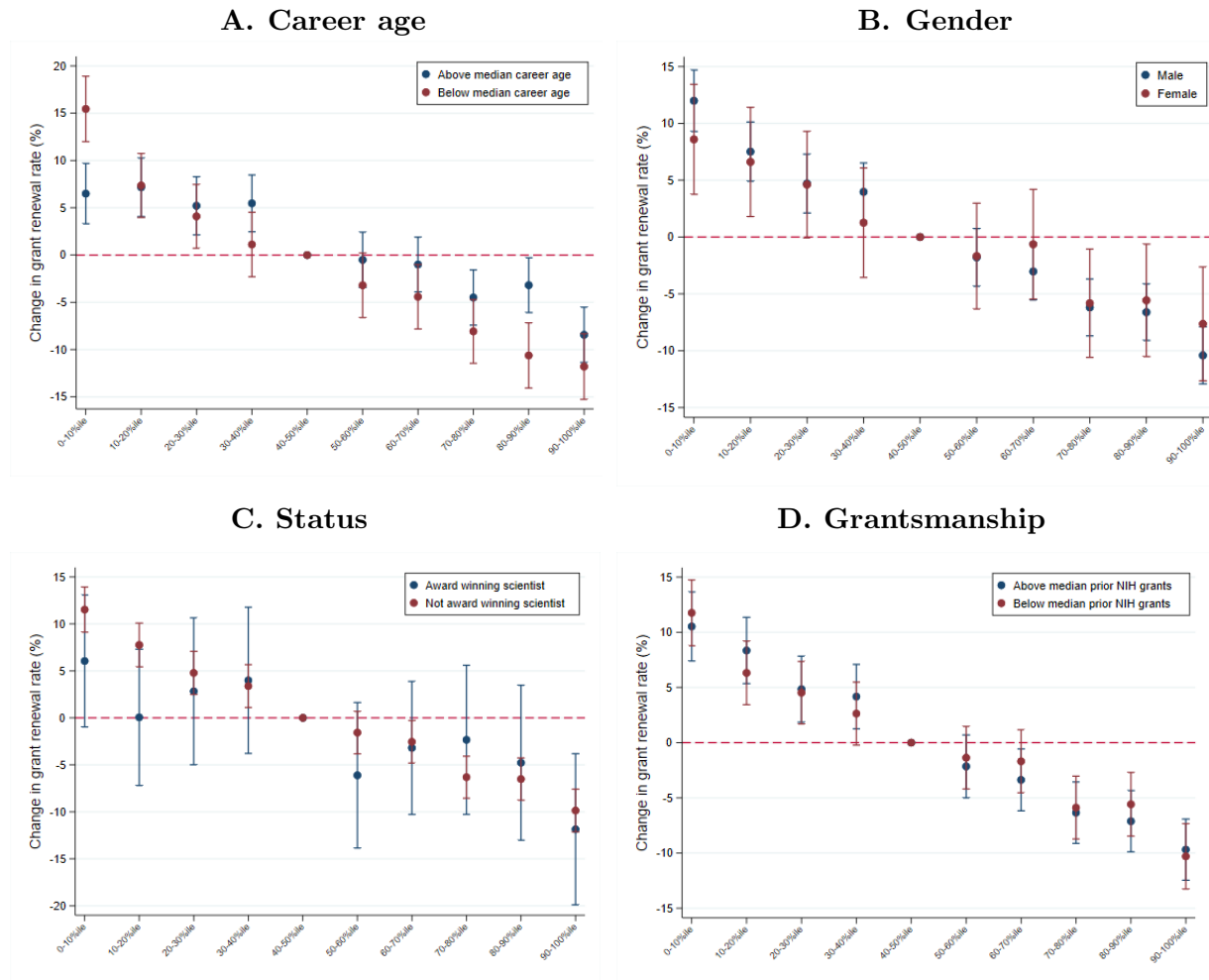
Note: The plots above present the coefficient estimates (and associated 95% confidence intervals) from a linear probability model where the dependent variable is an indicator variable for if the grant cycle was renewed. The unit of analysis is the investigator-grant-cycle. Panel A was modeled on Table 2, column 2, while panel B was modeled on Table 2, column 3. Each regression includes a set of 19 indicator variables for each ventile of the independent variable of interest, the fraction of MeSH term pairs from grant cycle publications that were not used by funded grant proposal abstracts in the same NIH IC-study section-year. The omitted category is the 45-50%ile. Robust standard errors were used, clustered at the investigator and grant.

Figure 7. Effect of risk taking and novelty on grant renewal



Note: The plots above present the coefficient estimates (and associated 95% confidence intervals) from a linear probability model where the dependent variable is an indicator variable for if the grant cycle was renewed modeled after Table 2, column 2. The unit of analysis is the investigator-grant-cycle. Each regression includes a set of 19 indicator variables for each ventile of the risk-taking measure, and 19 indicator variables for each ventile of novelty. The omitted category is the 45-50%ile. Robust standard errors were used, clustered at the investigator and grant

Figure 8. Heterogeneity in penalty for risk taking



Note: The plots above present the coefficient estimates (and associated 95% confidence intervals) from a linear probability model where the dependent variable is an indicator variable for if the grant cycle was renewed modeled after Table 2, column 2. The unit of analysis is the investigator-grant-cycle. In all cases the risk-taking measure is extreme tail outcomes. Each regression includes a set of 9 indicator variables for each decile of risk taking interacted with an indicator for above median career age and another 9 indicator variables for each decile of risk taking interacted with an indicator for below median career age (Panel A). Panel B similarly interacts on gender, panel C on elite status as measured by prominent scientific awards, and panel D on NIH funding prior to the grant cycle. The omitted category is the 40-50%ile. Robust standard errors were used, clustered at the investigator and grant.

Table 1a. Descriptive statistics: Investigator characteristics

	Mean	Median	Std. Dev.	Min.	Max.
Female	0.25	0	0.44	0	1
Degree year	1982.15	1982	12.26	1932	2012
Degree: PhD	0.68	1	0.47	0	1
Degree: MD	0.20	0	0.40	0	1
Degree: MD/PhD	0.10	0	0.31	0	1
Degree: Other	0.02	0	0.13	0	1
Nb of R01-equivalent grants	1.72	1	1.16	1	14
Nb of R01-equivalent grant cycles	3.05	2	2.78	1	37
Career NIH R01 funding (\$2020)	10,268,954	6,953,402	10,515,349	133,632	134,399,904
Career NIH funding (\$2020)	16,717,503	8,803,164	27,005,356	133,632	1,743,186,688
Career elite status	0.05	0	0.22	0	1

Note: Number of grants and cycles refers to only those grants that started 1980-2015. Career elite status identifies investigators who were ever won a Nobel Prize, Lasker Award, Howard Hughes Medical Institute Investigatorship, or membership in the Institute of Medicine or National Academy of Sciences. N=37,222 investigators.

Table 1b. Descriptive statistics: Grant cycle

	Renewed grant cycles					Not renewed grant cycles				
	Mean	Median	Std. Dev.	Min.	Max.	Mean	Median	Std. Dev.	Min.	Max.
Grant cycle number	1.94	1	1.28	1	10	1.69	1	1.15	1	11
First cycle of grant	0.51	1	0.50	0	1	0.62	1	0.49	0	1
Year started	1997.85	1999	9.16	1980	2015	2000.38	2002	9.82	1980	2015
Cycle length (years)	4.37	4	1.59	1	19	3.16	3	1.11	1	16
Time, degree to cycle start (years)	17.62	16	8.48	-22	61	19.57	18	9.42	-18	66
Investigator pubs. prior to cycle	53.35	36	57.52	0	1148	58.66	39	63.74	0	1196
Investigator citations prior to cycle	2,118.12	887	3,952.72	0	89,857	2,329.03	967	4,443.51	0	184,460
Publications acknowledging grant										
Nb of publications	8.06	6	7.87	1	184	7.01	5	7.34	1	164
Journal impact factor	6.28	5	4.00	0	72	5.36	5	3.51	0	43
Citations	662.91	355	1,146.86	1	6,4421	421.90	201	797.27	1	67,451
Highly selective journal	0.12	0	0.33	0	1	0.07	0	0.25	0	1
Publications not acknowledging grant										
Nb of publications	9.38	5	13.25	0	328	11.33	6	15.26	0	258
Journal impact factor	6.08	5	4.25	0	43	5.19	4	3.42	0	43
Citations	824.09	291	1,764.14	0	50,927	780.04	287	1,669.59	0	65,447
Highly selective journal	0.17	0	0.37	0	1	0.12	0	0.33	0	1
Fraction with any NIH funding	0.58	1	0.33	0	1	0.60	1	0.32	0	1

Note: Unit of analysis is at the investigator-grant-cycle level. N=103,164 (46,619 renewed and 56,545 not renewed investigator-grant-cycles). Grant cycle number minus one is the number of time the grant has previously been competitively renewed. Time from degree to cycle is measured from the last terminal degree earned by the investigator; 55 individuals (92 investigator-grant-cycles) earned an R01-equivalent grant prior to their last degree. This largely reflects investigators with multiple doctorates, frequently with large time gaps between them. Publications acknowledging a grant reflect those publications citing funding from the focal grant which were published up to one year after the conclusion of the grant cycle; publications not acknowledging the focal grant represents the investigator's other publications during the same time period. Citations are measured through 2020.

Table 1c. Descriptive statistics: Measures of risk taking and novelty

	Renewed grant cycles					Not renewed grant cycles				
	Mean	Median	Std. Dev.	Min.	Max.	Mean	Median	Std. Dev.	Min.	Max.
Extreme tail outcomes	46.69	47	23.80	0	100	47.40	48	23.79	0	100
Disruption	62.43	63	21.35	0	100	64.59	66	20.77	0	100
Pivoting	0.73	1	0.14	0	1	0.74	1	0.14	0	1
Standing out	0.87	1	0.11	0	1	0.85	1	0.11	0	1
Novelty	0.08	0	0.06	0	1	0.07	0	0.06	0	1

Note: Outcomes are measured at the investigator-grant-cycle level (N=103,164 investigator-grant-cycles). Measures are only defined for those cycles with qualifying associated publications (see footnote 7 for details).

Table 1d. Descriptive statistics: Correlations across measures of risk taking and novelty

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)
(1) Extreme tail outcomes	1.000						
(2) Disruption	0.399	1.000					
(3) Pivoting	-0.062	0.029	1.000				
(4) Standing out	0.099	0.036	0.118	1.000			
(5) Novelty	-0.055	-0.030	0.317	0.053	1.000		
(6) Nb. of publications	0.508	0.351	-0.125	0.119	-0.021	1.000	
(7) Log citations	0.353	0.192	-0.026	0.131	0.103	0.613	1.000

Note: N=103,164 investigator-grant-cycles. Measures are only defined for those cycles with qualifying associated publications (see footnote 7 for details). Number of publications and citations are based solely on publications acknowledging funding from the focal grant during the grant cycle. Citations are measured through 2020.

Table 2. Effect of risk taking on grant renewal

A. Extreme tail outcomes

	All renewals			First renewal			Second and future renewals		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
<10%ile, max-min citation percentile	0.054** (0.005)	0.051** (0.004)	0.022** (0.005)	0.061** (0.005)	0.055** (0.005)	0.028** (0.006)	0.064** (0.008)	0.057** (0.008)	0.018* (0.008)
>90%ile, max-min citation percentile	-0.039** (0.004)	-0.044** (0.004)	-0.021** (0.004)	-0.039** (0.006)	-0.040** (0.006)	-0.018** (0.006)	-0.047** (0.006)	-0.048** (0.006)	-0.020** (0.006)
Investigator demographics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Grant year and length	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
IC x study section	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Institutional quality	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Investigator quality	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Pub. quantity, quality during cycle	No	No	Yes	No	No	Yes	No	No	Yes
Mean of dependent variable	0.463	0.463	0.463	0.413	0.413	0.413	0.523	0.523	0.523
Std. Dev. of dependent variable	0.499	0.499	0.499	0.492	0.492	0.492	0.499	0.499	0.499
Effect bottom %ile group, in s.d. units	0.108	0.103	0.044	0.123	0.111	0.057	0.129	0.115	0.037
Effect top %ile group, in s.d. units	-0.078	-0.088	-0.041	-0.080	-0.082	-0.037	-0.093	-0.095	-0.040
Adjusted R ²	0.4685	0.5026	0.5175	0.5055	0.5267	0.5370	0.4572	0.4983	0.5166
Nb. of investigators	33,409	33,403	33,403	30,446	30,445	30,445	17,274	17,260	17,260
Nb. of investigator-grant-cycles	90,123	90,075	90,075	48,530	48,529	48,529	40,859	40,810	40,810

Note: Estimates stem from a linear probability model where the unit of analysis is the investigator-grant-cycle and the dependent variable is an indicator variable for if the grant cycle was renewed. Columns 1-3 includes the full sample, columns 4-6 limit the sample to the first cycle for each investigator-grant, and columns 7-9 limit it to cycles 2 and above for each investigator-grant. The independent variables are indicator variables for <10%ile and >90%ile for the difference between the maximum and minimum vintage-adjusted citation percentile of grant cycle publications, with the 10-90%iles as the omitted category. Columns 1, 4 and 7 include indicator variables for gender, investigator degree, investigator career age, year of grant evaluation, length of the grant cycle, and a set of 2,911 indicator variables corresponding to each NIH institute or center and study section. Columns 2, 5, and 8 include all of the previous controls, but also include log NIH funding to the investigator's institution, log investigator prior career citations, log investigator prior NIH grant funding, and an indicator for no prior NIH grant funding. Finally, columns 3, 6, and 9 include all previous controls, but also add log mean journal impact factor, indicator variables for grant and non-grant publications in one of four highly selective journals, and a set of 21 indicator variables separately for the number of publications during the grant cycle acknowledging and not acknowledging the grant. Robust standard errors were used, clustering at the investigator and grant. [†] $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

B. Disruption

	All renewals			First renewal			Second and future renewals		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
<10%ile, maximum disruption index	0.091** (0.012)	0.089** (0.012)	0.054** (0.012)	0.094** (0.014)	0.089** (0.013)	0.057** (0.013)	0.080** (0.024)	0.078** (0.023)	0.037 (0.023)
>90%ile, maximum disruption index	-0.019** (0.004)	-0.022** (0.004)	0.008† (0.004)	-0.020** (0.005)	-0.020** (0.005)	0.008 (0.005)	-0.023** (0.006)	-0.024** (0.006)	0.009 (0.006)
Investigator demographics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Grant year and length	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
IC x study section	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Institutional quality	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Investigator quality	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Pub. quantity, quality during cycle	No	No	Yes	No	No	Yes	No	No	Yes
Mean of dependent variable	0.454	0.454	0.454	0.405	0.405	0.405	0.519	0.519	0.519
Std. Dev. of dependent variable	0.498	0.498	0.498	0.491	0.491	0.491	0.500	0.500	0.500
Effect bottom %ile group, in s.d. units	0.183	0.179	0.108	0.192	0.180	0.116	0.159	0.156	0.075
Effect top %ile group, in s.d. units	-0.039	-0.044	0.016	-0.040	-0.041	0.016	-0.045	-0.048	0.017
Adjusted R ²	0.4481	0.4833	0.5017	0.4798	0.5028	0.5169	0.4427	0.4861	0.5074
Nb. of investigators	36,876	36,876	36,876	34,889	34,889	34,889	18,196	18,196	18,196
Nb. of investigator-grant-cycles	102,250	102,250	102,250	57,667	57,667	57,667	43,832	43,832	43,832

Note: Estimates stem from a linear probability model where the unit of analysis is the investigator-grant-cycle and the dependent variable is an indicator variable for if the grant cycle was renewed. Columns 1-3 includes the full sample, columns 4-6 limit the sample to the first cycle for each investigator-grant, and columns 7-9 limit it to cycles 2 and above for each investigator-grant. The independent variables are indicator variables for <10%ile and >90%ile for the maximum of the disruption index for grant cycle publications, with the 10-90%iles as the omitted category. Columns 1, 4 and 7 include indicator variables for gender, investigator degree, investigator career age, year of grant evaluation, length of the grant cycle, and a set of 2,911 indicator variables corresponding to each NIH institute or center and study section. Columns 2, 5, and 8 include all of the previous controls, but also include log NIH funding to the investigator's institution, log investigator prior career citations, log investigator prior NIH grant funding, and an indicator for no prior NIH grant funding. Finally, columns 3, 6, and 9 include all previous controls, but also add log mean journal impact factor, indicator variables for grant and non-grant publications in one of four highly selective journals, and a set of 21 indicator variables separately for the number of publications during the grant cycle acknowledging and not acknowledging the grant. Robust standard errors were used, clustering at the investigator and grant. † $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

C. Pivoting

	All renewals			First renewal			Second and future renewals		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
<10%ile, fraction new MeSH pairs	0.068** (0.004)	0.053** (0.004)	0.059** (0.004)	0.052** (0.006)	0.042** (0.006)	0.044** (0.006)	0.071** (0.006)	0.057** (0.006)	0.063** (0.006)
>90%ile, fraction new MeSH pairs	-0.039** (0.004)	-0.024** (0.004)	-0.043** (0.004)	-0.019** (0.004)	-0.012** (0.004)	-0.030** (0.004)	-0.052** (0.009)	-0.039** (0.008)	-0.062** (0.008)
Investigator demographics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Grant year and length	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
IC x study section	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Institutional quality	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Investigator quality	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Pub. quantity, quality during cycle	No	No	Yes	No	No	Yes	No	No	Yes
Mean of dependent variable	0.455	0.455	0.455	0.406	0.406	0.406	0.519	0.519	0.519
Std. Dev. of dependent variable	0.498	0.498	0.498	0.491	0.491	0.491	0.500	0.500	0.500
Effect bottom %ile group, in s.d. units	0.137	0.107	0.119	0.106	0.085	0.090	0.141	0.115	0.127
Effect top %ile group, in s.d. units	-0.079	-0.048	-0.086	-0.040	-0.025	-0.060	-0.105	-0.077	-0.124
Adjusted R ²	0.4499	0.4847	0.5039	0.4802	0.5032	0.5181	0.4453	0.4887	0.5103
Nb. of investigators	36,544	36,539	36,539	34,497	34,497	34,497	18,127	18,114	18,114
Nb. of investigator-grant-cycles	101,389	101,338	101,338	57,008	57,007	57,007	43,629	43,577	43,577

Note: Estimates stem from a linear probability model where the unit of analysis is the investigator-grant-cycle and the dependent variable is an indicator variable for if the grant cycle was renewed. Columns 1-3 includes the full sample, columns 4-6 limit the sample to the first cycle for each investigator-grant, and columns 7-9 limit it to cycles 2 and above for each investigator-grant. The independent variables are indicator variables for <10%ile and >90%ile for the fraction of MeSH term pairs from grant cycle publications that were not used by the investigator in the 5 years preceding the grant cycle, with the 10-90%iles as the omitted category. Columns 1, 4 and 7 include indicator variables for gender, investigator degree, investigator career age, year of grant evaluation, length of the grant cycle, and a set of 2,911 indicator variables corresponding to each NIH institute or center and study section. Columns 2, 5, and 8 include all of the previous controls, but also include log NIH funding to the investigator's institution, log investigator prior career citations, log investigator prior NIH grant funding, and an indicator for no prior NIH grant funding. Finally, columns 3, 6, and 9 include all previous controls, but also add log mean journal impact factor, indicator variables for grant and non-grant publications in one of four highly selective journals, and a set of 21 indicator variables separately for the number of publications during the grant cycle acknowledging and not acknowledging the grant. Robust standard errors were used, clustering at the investigator and grant. [†] $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

D. Standing out

	All renewals			First renewal			Second and future renewals		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
<10%ile, fraction unique MeSH pairs	0.040** (0.005)	0.038** (0.005)	0.025** (0.005)	0.037** (0.006)	0.035** (0.006)	0.022** (0.006)	0.048** (0.009)	0.044** (0.008)	0.031** (0.008)
>90%ile, fraction unique MeSH pairs	-0.022** (0.005)	-0.023** (0.005)	-0.027** (0.005)	-0.003 (0.007)	-0.003 (0.007)	-0.010 (0.006)	-0.035** (0.007)	-0.041** (0.007)	-0.041** (0.007)
Investigator demographics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Grant year and length	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
IC x study section	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Institutional quality	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Investigator quality	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Pub. quantity, quality during cycle	No	No	Yes	No	No	Yes	No	No	Yes
Mean of dependent variable	0.491	0.491	0.491	0.443	0.443	0.443	0.551	0.551	0.551
Std. Dev. of dependent variable	0.500	0.500	0.500	0.497	0.497	0.497	0.497	0.497	0.497
Effect bottom %ile group, in s.d. units	0.081	0.077	0.050	0.075	0.070	0.044	0.096	0.088	0.061
Effect top %ile group, in s.d. units	-0.043	-0.046	-0.053	-0.006	-0.007	-0.020	-0.071	-0.083	-0.083
Adjusted R ²	0.4294	0.4688	0.4905	0.4690	0.4941	0.5106	0.4200	0.4694	0.4946
Nb. of investigators	32,370	32,364	32,364	30,209	30,209	30,209	16,041	16,031	16,031
Nb. of investigator-grant-cycles	83,005	82,962	82,962	46,902	46,901	46,901	35,650	35,609	35,609

Note: Estimates stem from a linear probability model where the unit of analysis is the investigator-grant-cycle and the dependent variable is an indicator variable for if the grant cycle was renewed. Columns 1-3 includes the full sample, columns 4-6 limit the sample to the first cycle for each investigator-grant, and columns 7-9 limit it to cycles 2 and above for each investigator-grant. The independent variables are indicator variables for <10%ile and >90%ile for the fraction of MeSH term pairs from grant cycle publications that were not used by funded grant proposal abstracts in the same NIH IC-study section-year, with the 10-90%iles as the omitted category. Columns 1, 4 and 7 include indicator variables for gender, investigator degree, investigator career age, year of grant evaluation, length of the grant cycle, and a set of 2,911 indicator variables corresponding to each NIH institute or center and study section. Columns 2, 5, and 8 include all of the previous controls, but also include log NIH funding to the investigator's institution, log investigator prior career citations, log investigator prior NIH grant funding, and an indicator for no prior NIH grant funding. Finally, columns 3, 6, and 9 include all previous controls, but also add log mean journal impact factor, indicator variables for grant and non-grant publications in one of four highly selective journals, and a set of 21 indicator variables separately for the number of publications during the grant cycle acknowledging and not acknowledging the grant. Robust standard errors were used, clustering at the investigator and grant. [†] $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

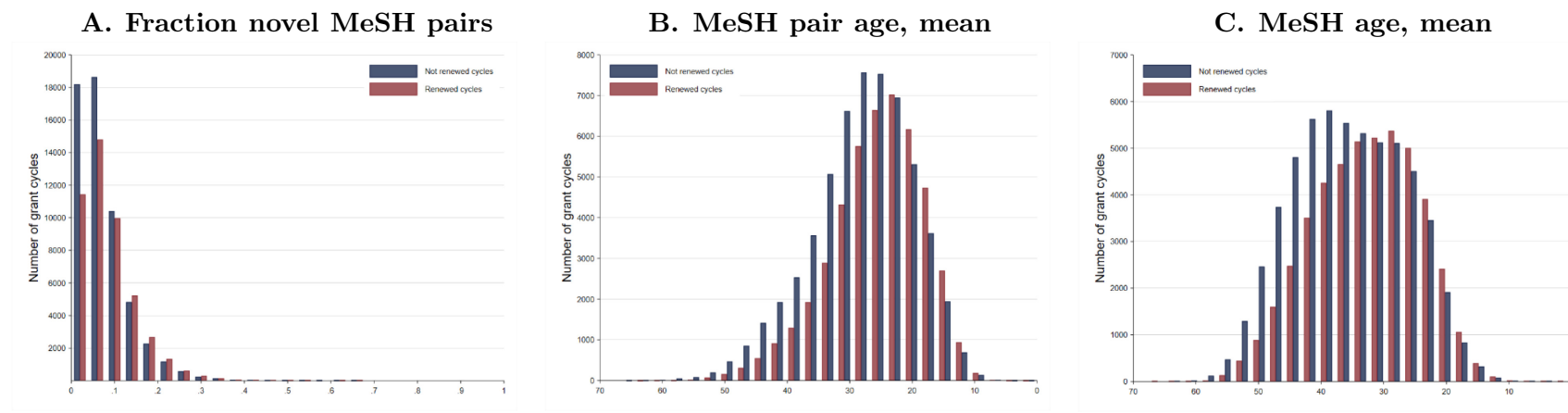
Table 3. Effect of novelty on the risk taking grant renewal penalty

	Base controls				Grant output controls			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Novelty >50%ile	0.0254** (0.0057)	0.0211** (0.0077)	0.0922** (0.0142)	0.1137** (0.0230)	0.0284** (0.0056)	0.0191* (0.0076)	0.1200** (0.0139)	0.1252** (0.0225)
Extreme tail outcomes	-0.0117** (0.0008)				-0.0058** (0.0009)			
Extreme tail outcomes × Novelty >50%ile	-0.0025* (0.0010)				-0.0025* (0.0010)			
Disruption		-0.0145** (0.0008)				-0.0078** (0.0008)		
Disruption × Novelty >50%ile		-0.0015 (0.0011)				-0.0007 (0.0011)		
Pivoting			-0.1601** (0.0121)				-0.2124** (0.0119)	
Pivoting × Novelty >50%ile			-0.0922** (0.0184)				-0.1178** (0.0181)	
Standing out				-0.1897** (0.0201)				-0.1513** (0.0198)
Standing out × Novelty >50%ile				-0.1135** (0.0261)				-0.1215** (0.0255)

Note: Estimates stem from a linear probability model where the unit of analysis is the investigator-grant-cycle and the dependent variable is an indicator variable for if the grant cycle was renewed. Columns with base controls are very similar to Table 2, column 2, while those with grant output controls are very similar to Table 2, column 3. Each estimate includes a risk-taking measure as a continuous variable: extreme tail outcomes scaled so a change of 1 in the regressor reflects a change of 10 in the long-run vintage-adjusted citation percentile spread (columns 1 and 5), disruption scaled so a change of 1 in the regressor reflects a change of 10 in the maximum disruption index percentile (columns 2 and 6), pivoting (columns 3 and 7), and standing out (columns 4 and 8). All estimates include an indicator variable for above median novelty as well as an interaction term between this indicator variable and the risk-taking measure. Robust standard errors were used, clustered at the investigator and grant. [†] $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

Appendix A: Ancillary Results and Robustness Checks

Figure A1. Distribution of measures of novelty

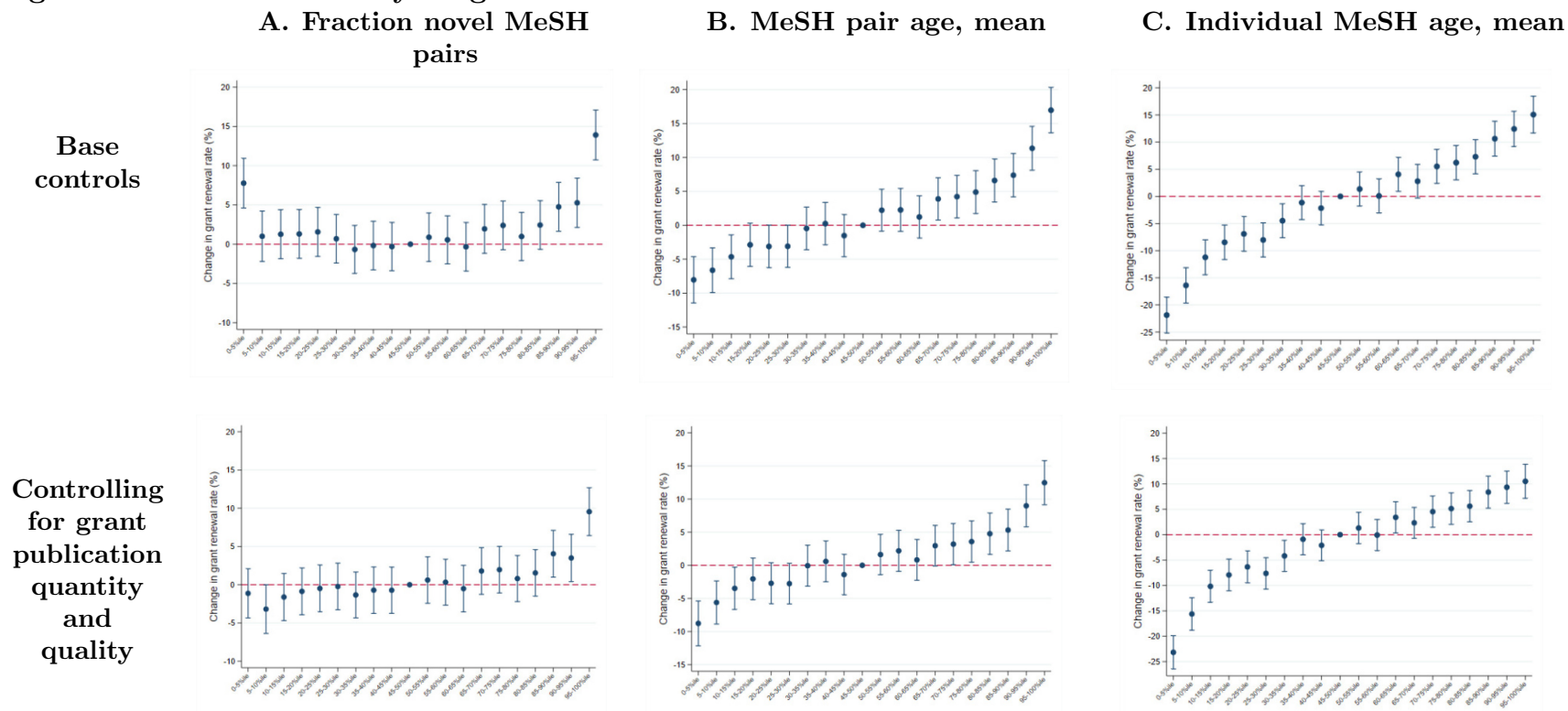


Note: Histogram of the fraction of MeSH pairs among grant cycle publications that were first used by any PubMed publication within the prior 3 years. Unit of analysis is at the investigator-grant-cycle level.

Note: Histogram of the mean MeSH pair age among all grant cycle publication MeSH pairs. Age is defined as the difference between the year of the focal publication and the year the MeSH pair was first used by any publication in PubMed. Unit of analysis is at the investigator-grant-cycle level.

Note: Histogram of the mean MeSH age among all grant cycle publication MeSH. Age is defined as the difference between the year of the focal publication and the year the MeSH was first used by any publication in PubMed. Unit of analysis is at the investigator-grant-cycle level.

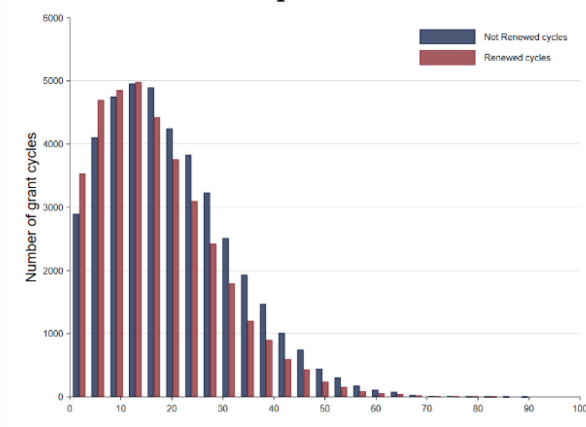
Figure A2. Effect of novelty on grant renewal



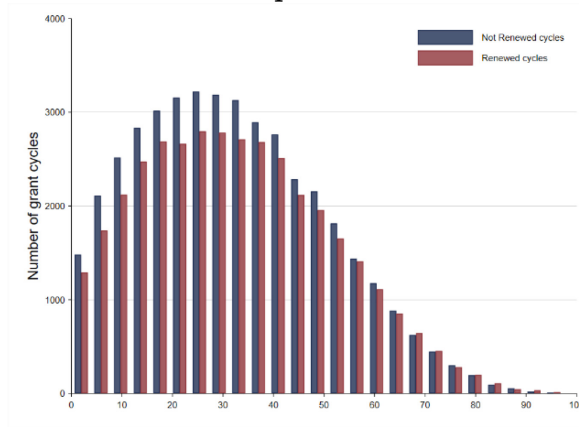
Note: The plots above present the coefficient estimates (and associated 95% confidence intervals) from a linear probability model where the dependent variable is an indicator variable for if the grant cycle was renewed. The unit of analysis is the investigator-grant-cycle. Row 1 was modeled on Table 2, column 2, while row 2 was modeled on Table 2, column 3. Each regression includes a set of 19 indicator variables for each ventile of the independent variable of interest, which is the fraction of MeSH pairs first used in PubMed within the prior 3 years (Panel A), the mean MeSH pair age (Panel B) and mean MeSH age (Panel C) among all MeSH and MeSH pairs for grant cycle publications. For ease of comparison to other figures, age is inverted so that higher percentiles represent grant cycles with younger mean ages. The omitted category is the 45-50%ile. Robust standard errors were used, clustered at the investigator and grant.

Figure A3. Distribution of alternative measures of extreme tail outcomes

A. Range, max-median citation percentile



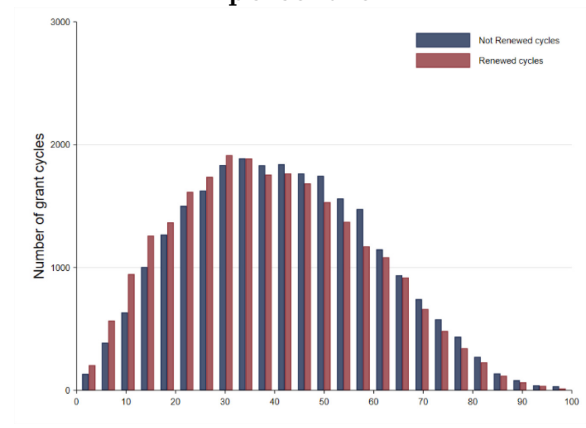
B. Range, median-min citation percentile



Note: Histogram of the range between the maximum and median vintage-adjusted citation percentile for grant cycle publications. Unit of analysis is at the investigator-grant-cycle level.

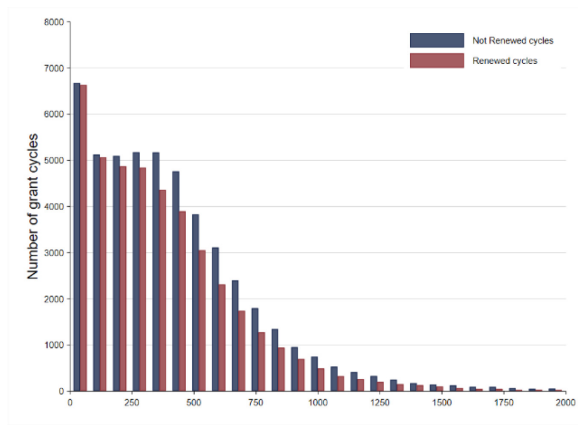
Note: Histogram of the range between the median and minimum vintage-adjusted citation percentile for grant cycle publications. Unit of analysis is at the investigator-grant-cycle level.

C. Range, winsorized max-min citation percentile



Note: Histogram of the range between the winsorized maximum and minimum vintage-adjusted citation percentile for grant cycle publications. Unit of analysis is at the investigator-grant-cycle level.

D. Variance, citation percentile



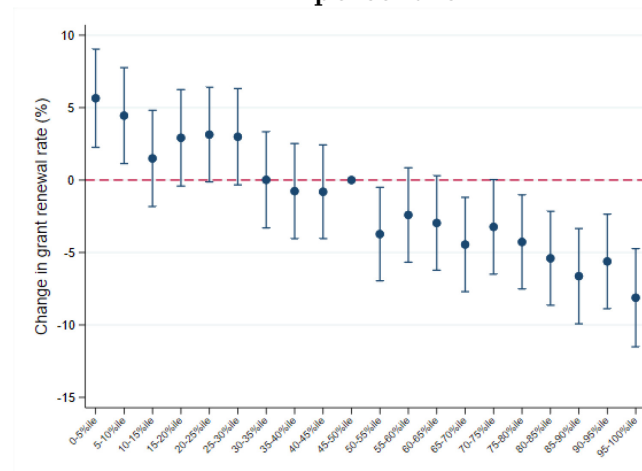
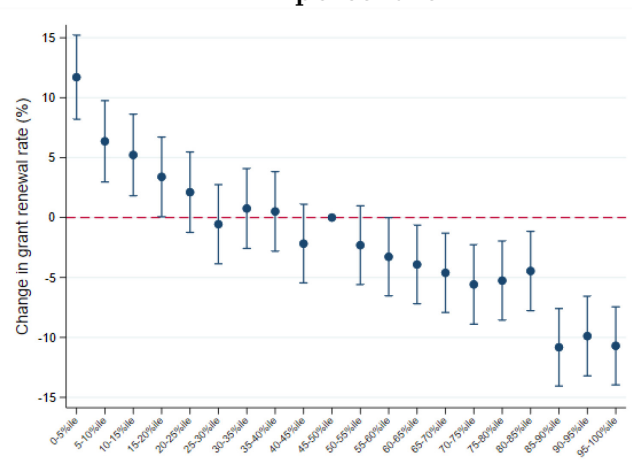
Note: Histogram of the variance in vintage-adjusted citation percentile for grant cycle publications. Unit of analysis is at the investigator-grant-cycle level. Four hundred seventy four outliers with a variance over 2,000 are not shown.

Figure A4. Effect of risk taking on grant renewal: Alternative measures of extreme tail outcomes

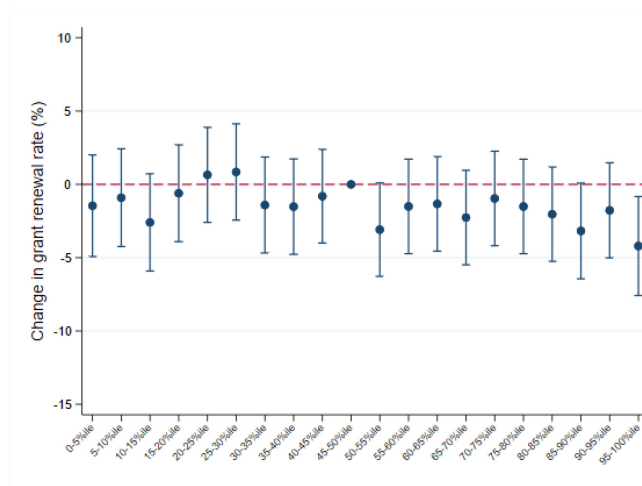
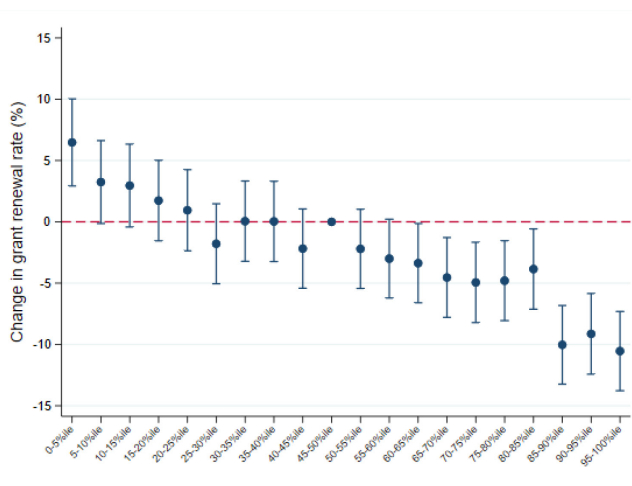
A. Range, max-median citation
percentile

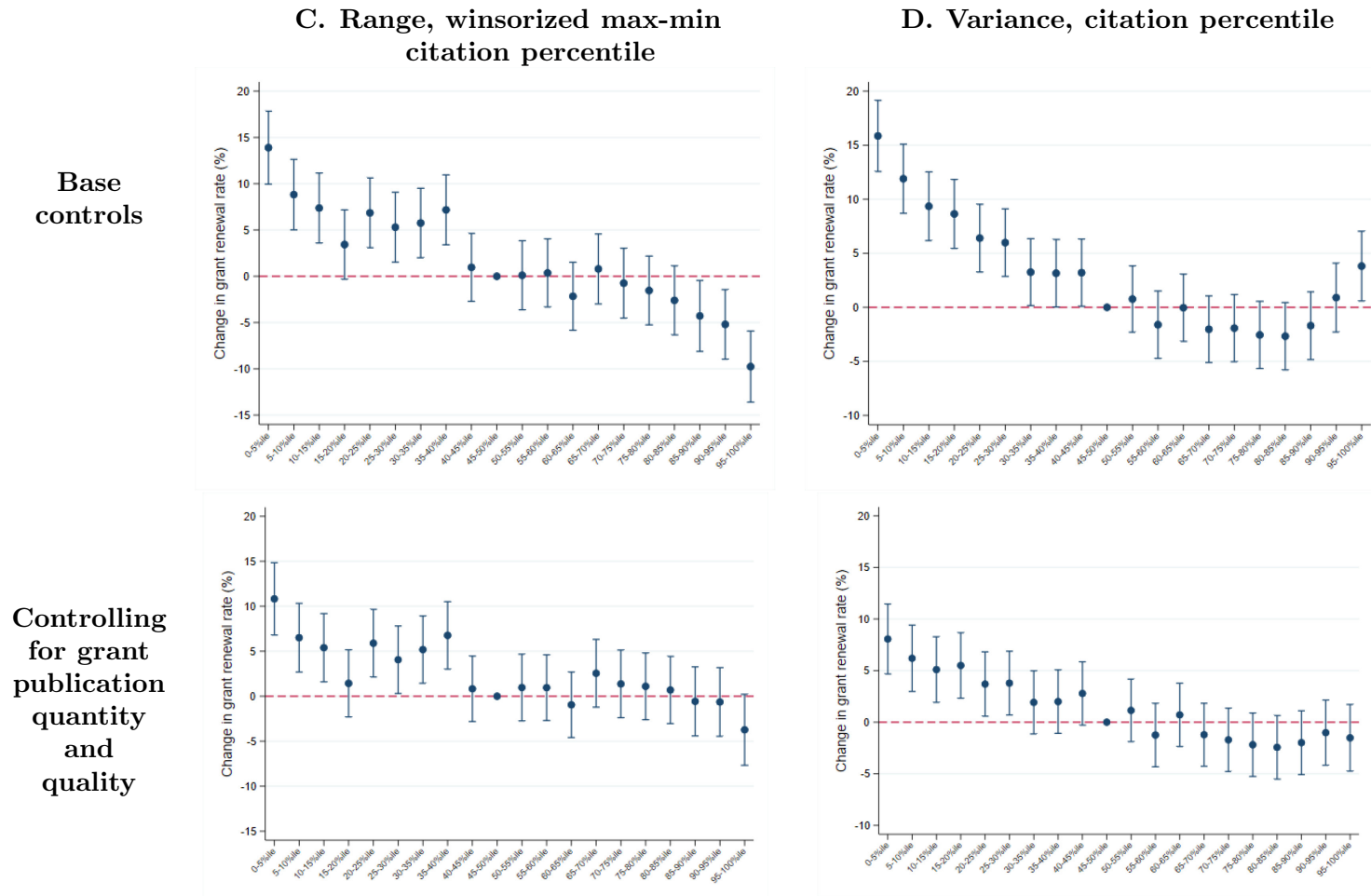
B. Range, median-min citation
percentile

Base
controls



Controlling
for grant
publication
quantity
and
quality

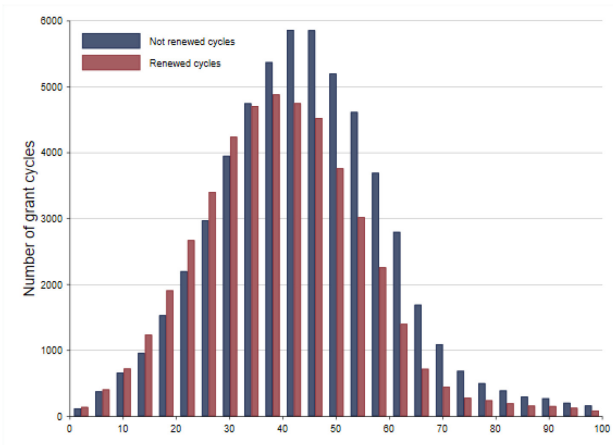




Note: The plots above present the coefficient estimates (and associated 95% confidence intervals) from a linear probability model where the dependent variable is an indicator variable for if the grant cycle was renewed. The unit of analysis is the investigator-grant-cycle. Row 1 was modeled on Table 2, column 2, while row 2 was modeled on Table 2, column 3. Each regression includes a set of 19 indicator variables for each ventile of the independent variable of interest: the difference between the maximum and median vintage-adjusted citation percentile (Panel A), difference between the median and minimum vintage-adjusted citation percentile (Panel B), difference between the winsorized maximum and minimum vintage-adjusted citation percentile (Panel C), and variance in vintage-adjusted citation percentile (Panel D) among grant cycle publications. The omitted category is the 45-50%ile. Robust standard errors were used, clustered at the investigator and grant.

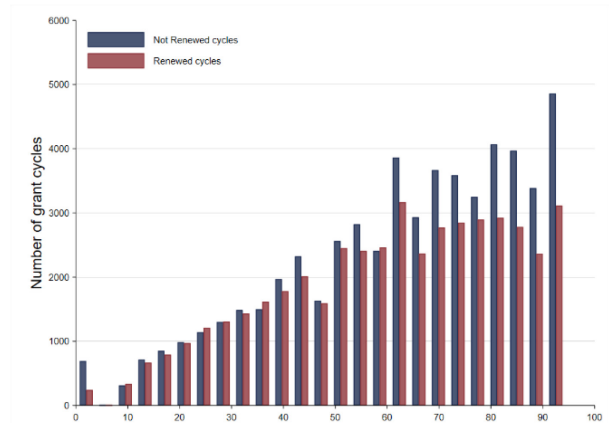
Figure A5. Distribution of alternative measures of disruption

A. Disruption index percentile, mean



Note: Histogram of the mean disruption index percentile among grant cycle publications. Unit of analysis is at the investigator-grant-cycle level.

B. Disruption index without Nk term percentile, maximum



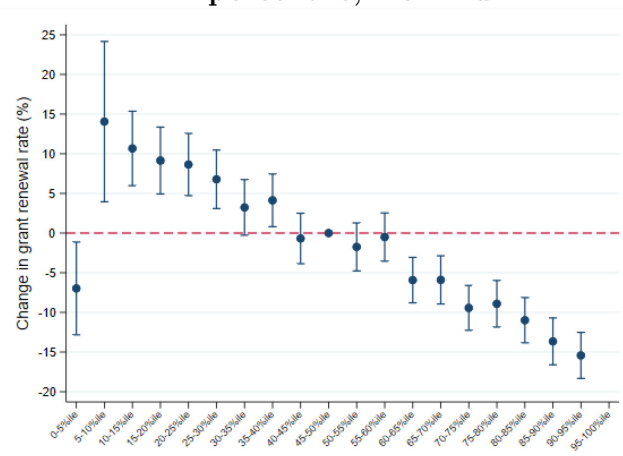
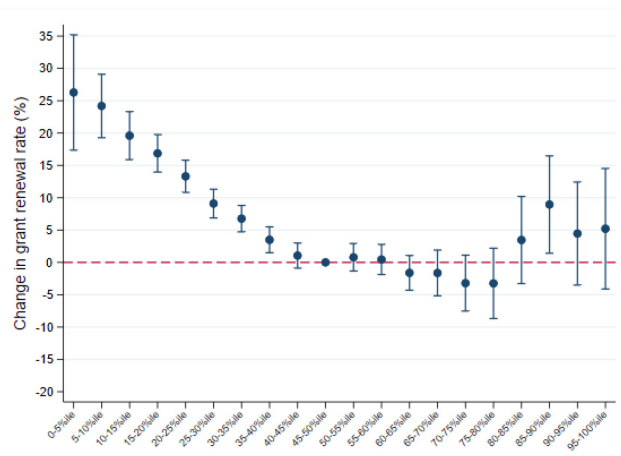
Note: Histogram of the maximum disruption index percentile among grant cycle publications. This alternative specification of the disruption index follows Bornmann et al. (2020) in excluding the Nk term, which adjusts the disruption index for citations to a focal paper's references that do not cite the focal paper itself. Unit of analysis is at the investigator-grant-cycle level.

Figure A6. Effect of risk taking on grant renewal: Alternative measures of disruption

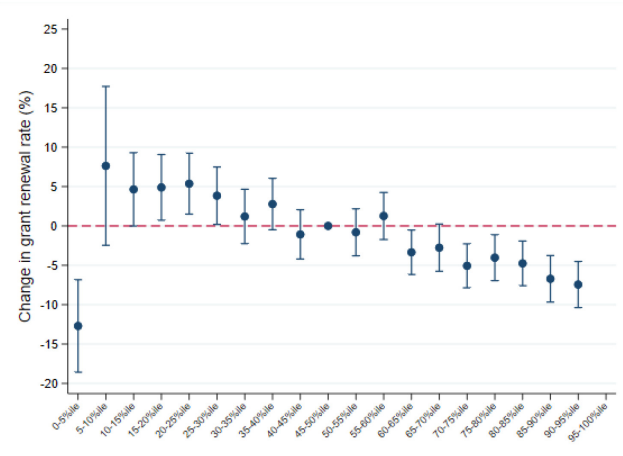
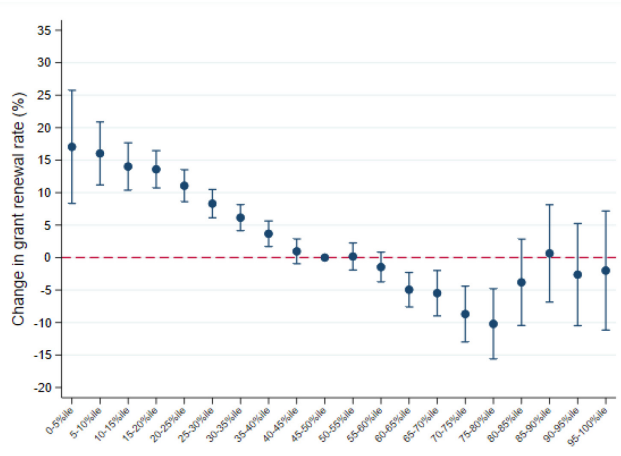
A. Disruption index percentile, mean

**B. Disruption index without Nk term
percentile, maximum**

**Base
controls**

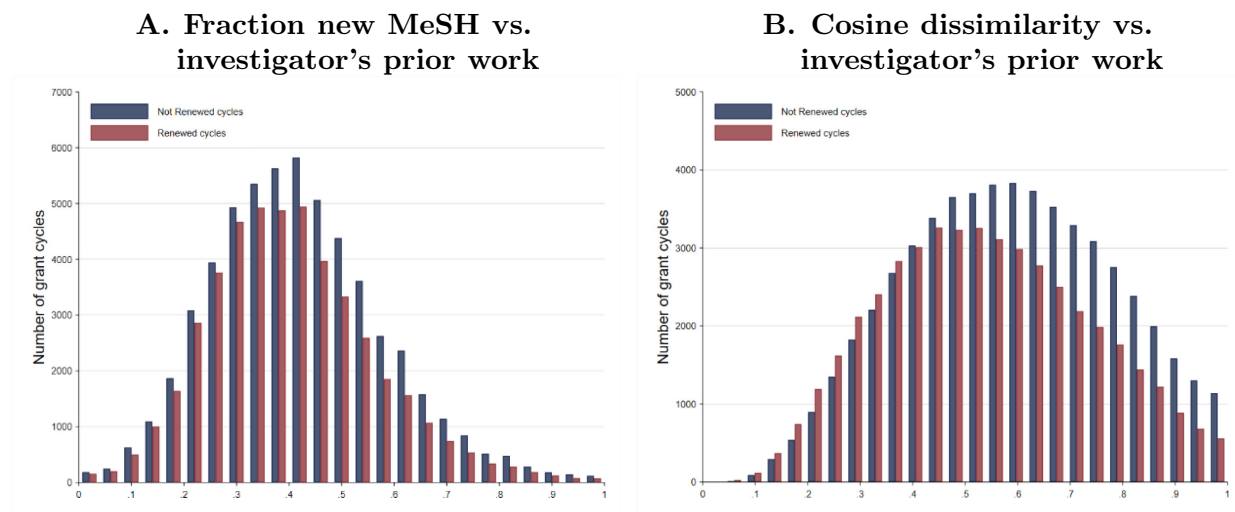


**Controlling
for grant
publication
quantity
and
quality**



Note: The plots above present the coefficient estimates (and associated 95% confidence intervals) from a linear probability model where the dependent variable is an indicator variable for if the grant cycle was renewed. The unit of analysis is the investigator-grant-cycle. Row 1 was modeled on Table 2, column 2, while row 2 was modeled on Table 2, column 3. Each regression includes a set of 19 indicator variables for each ventile of the independent variable of interest, which is the mean disruption index percentile (Panel A) and maximum disruption index percentile following Bormann et al. (2020) in excluding the Nk term (Panel B) among grant cycle publications. The omitted category is the 45-50%ile. Robust standard errors were used, clustered at the investigator and grant.

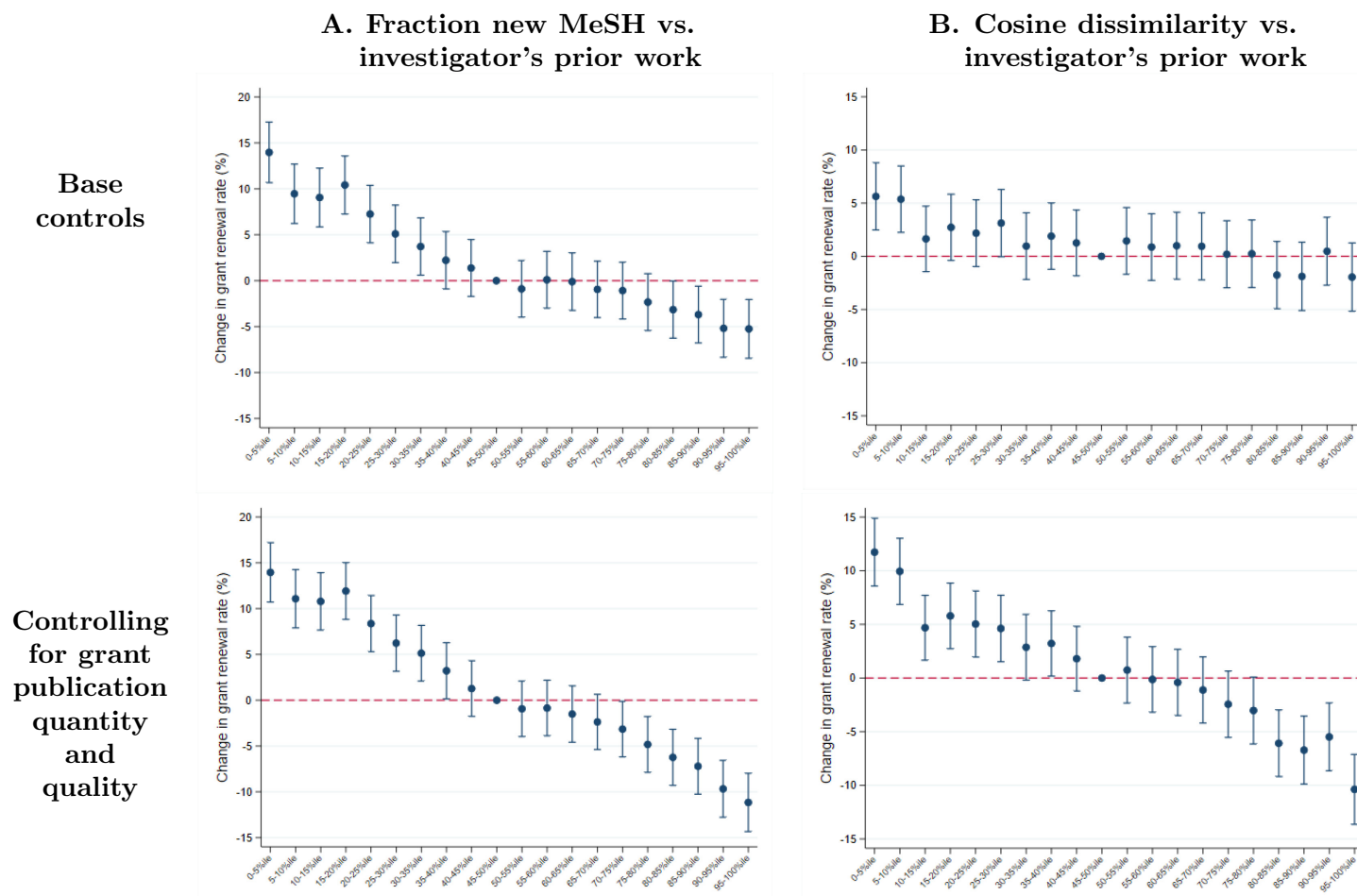
Figure A7. Distribution of alternative measures of pivoting



Note: Histogram of the fraction of MeSH terms from grant cycle publications that were not used in the MeSH for the investigator's publications in the 5 years preceding the grant cycle. Unit of analysis is at the investigator-grant-cycle level.

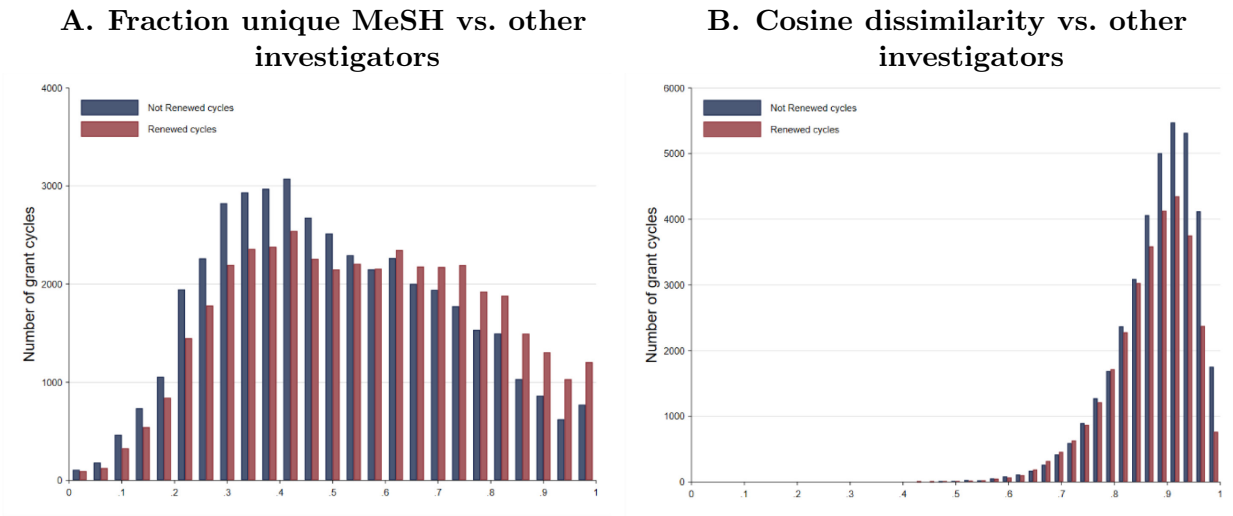
Note: Histogram of the cosine dissimilarity between the MeSH of grant cycle publications and MeSH for the investigator's publications in the 5 years preceding the grant. MeSH were weighted using term-frequency-inverse document frequency. Unit of analysis is at the investigator-grant-cycle level.

Figure A8. Effect of risk taking on grant renewal: Alternative measures of pivoting



Note: The plots above present the coefficient estimates (and associated 95% confidence intervals) from a linear probability model where the dependent variable is an indicator variable for if the grant cycle was renewed. The unit of analysis is the investigator-grant-cycle. Row 1 was modeled on Table 2, column 2, while row 2 was modeled on Table 2, column 3. Each regression includes a set of 19 indicator variables for each ventile of the independent variable of interest, which is the fraction of new MeSH (Panel A) and cosine dissimilarity with term frequency-inverse document frequency weighting (Panel B) comparing MeSH from grant cycle publications to those for the investigator in the 5 years preceding the grant cycle. The omitted category is the 45-50%ile. Robust standard errors were used, clustered at the investigator and grant.

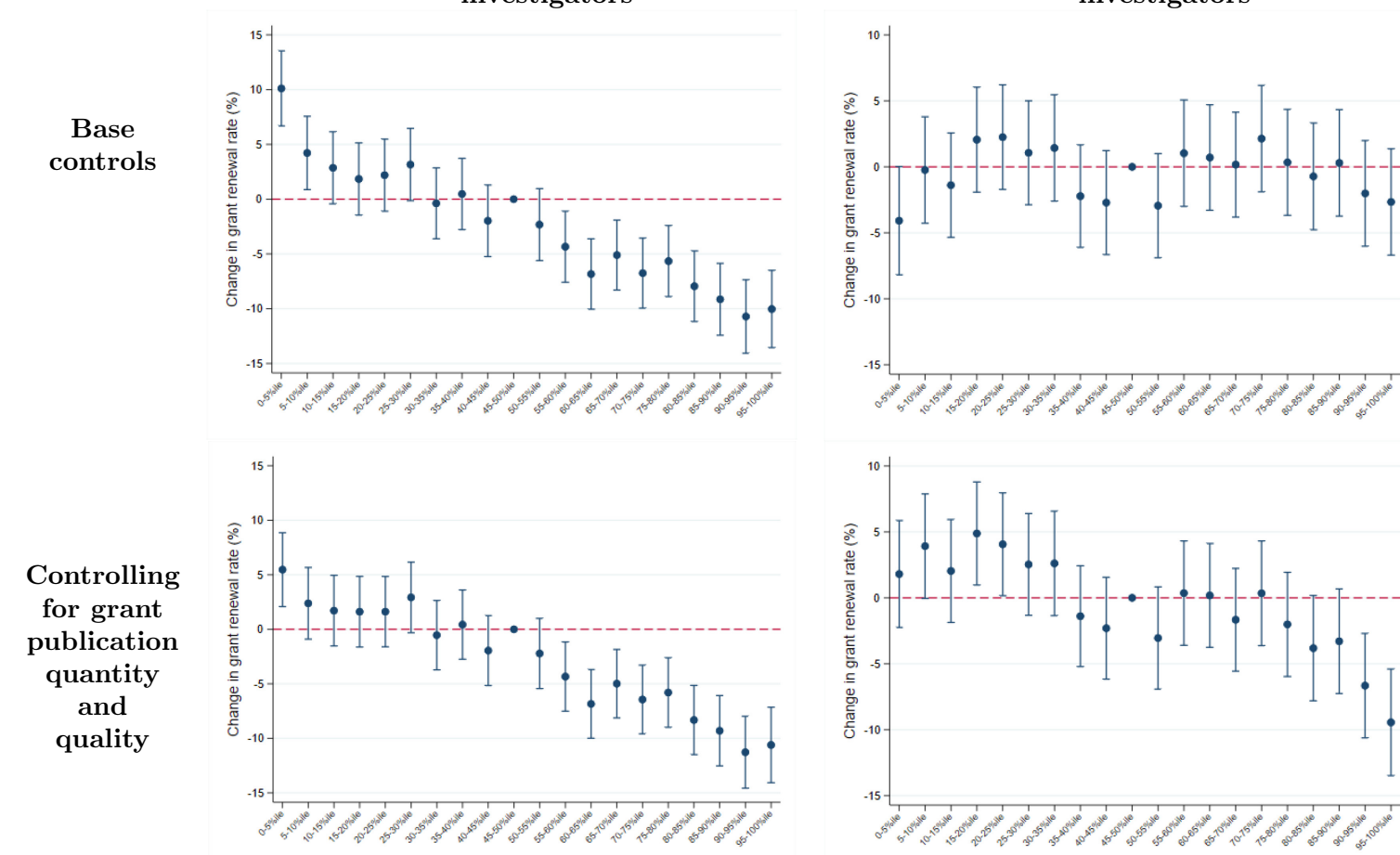
Figure A9. Distribution of alternative measures of standing out



Note: Histogram of the fraction of MeSH terms for grant cycle publications that were not used by any funded grant proposal abstracts in the same NIH IC-study section-year. Unit of analysis is at the investigator-grant-cycle level.

Note: Histogram of the cosine dissimilarity between the MeSH for grant cycle publications and all funded grant proposal abstracts in the same NIH IC-study section-year. MeSH were weighted using term-frequency-inverse document frequency. Unit of analysis is at the investigator-grant-cycle level.

Figure A10. Effect of risk taking on grant renewal: Alternative measures of standing out
A. Fraction unique MeSH vs. other investigators **B. Cosine dissimilarity vs. other investigators**



Note: The plots above present the coefficient estimates (and associated 95% confidence intervals) from a linear probability model where the dependent variable is an indicator variable for if the grant cycle was renewed. The unit of analysis is the investigator-grant-cycle. Row 1 was modeled on Table 2, column 2, while row 2 was modeled on Table 2, column 3. Each regression includes a set of 19 indicator variables for each ventile of the independent variable of interest, which is the fraction of unique MeSH (Panel A) and cosine dissimilarity with term frequency-inverse document frequency weighting (Panel B) comparing MeSH from grant cycle publications to those for funded grant proposal abstracts in the same NIH IC-study section-year. The omitted category is the 45-50%ile. Robust standard errors were used, clustered at the investigator and grant.

Table A1. Effect of novelty on the risk taking grant renewal penalty

A. Novelty above the 75%ile

	Base controls				Grant output controls			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Novelty >75%ile	0.0204** (0.0064)	0.0117 (0.0088)	0.1016** (0.0185)	0.1666** (0.0293)	0.0233** (0.0063)	0.0095 (0.0087)	0.1203** (0.0182)	0.1637** (0.0286)
Extreme tail outcomes	-0.0126** (0.0006)				-0.0069** (0.0007)			
Extreme tail outcomes × Novelty >75%ile	-0.0004 (0.0012)				-0.0007 (0.0012)			
Disruption		-0.0153** (0.0007)				-0.0084** (0.0007)		
Disruption × Novelty >75%ile		0.0010 (0.0013)				0.0014 (0.0013)		
Pivoting			-0.1807** (0.0106)				-0.2305** (0.0106)	
Pivoting × Novelty >75%ile			-0.0882** (0.0233)				-0.1107** (0.0229)	
Standing out				-0.2052** (0.0186)				-0.1660** (0.0183)
Standing out × Novelty >75%ile				-0.1593** (0.0328)				-0.1580** (0.0321)

Note: Estimates stem from a linear probability model where the unit of analysis is the investigator-grant-cycle and the dependent variable is an indicator variable for if the grant cycle was renewed. Columns with base controls are very similar to Table 2, column 2, while those with grant output controls are very similar to Table 2, column 3. Each estimate includes a risk-taking measure as a continuous variable: extreme tail outcomes scaled so a change of 1 in the regressor reflects a change of 10 in the long-run vintage-adjusted citation percentile spread (columns 1 and 5), disruption scaled so a change of 1 in the regressor reflects a change of 10 in the maximum disruption index percentile (columns 2 and 6), pivoting (columns 3 and 7), and standing out (columns 4 and 8). All estimates include an indicator variable for above the 75%ile in novelty as well as an interaction term between this indicator variable and the risk-taking measure. Robust standard errors were used, clustered at the investigator and grant. [†] $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

B. Novelty above the 90%ile

	Base controls				Grant output controls			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Novelty >90%ile	0.0206*	0.0167	0.1327**	0.2259**	0.0246**	0.0131	0.1359**	0.2085**
	(0.0090)	(0.0120)	(0.0284)	(0.0440)	(0.0089)	(0.0118)	(0.0279)	(0.0426)
Extreme tail outcomes	-0.0128**				-0.0072**			
	(0.0006)				(0.0007)			
Extreme tail outcomes × Novelty >90%ile	0.0018				0.0009			
	(0.0018)				(0.0017)			
Disruption		-0.0151**				-0.0083**		
		(0.0006)				(0.0007)		
Disruption × Novelty >90%ile		0.0024				0.0025		
		(0.0019)				(0.0018)		
Pivoting			-0.1855**				-0.2349**	
			(0.0101)				(0.0100)	
Pivoting × Novelty >90%ile			-0.1017**				-0.1136**	
			(0.0347)				(0.0342)	
Standing out				-0.2173**				-0.1777**
				(0.0181)				(0.0178)
Standing out × Novelty >90%ile				-0.2009**				-0.1911**
				(0.0488)				(0.0473)

Note: Estimates stem from a linear probability model where the unit of analysis is the investigator-grant-cycle and the dependent variable is an indicator variable for if the grant cycle was renewed. Columns with base controls are very similar to Table 2, column 2, while those with grant output controls are very similar to Table 2, column 3. Each estimate includes a risk-taking measure as a continuous variable: extreme tail outcomes scaled so a change of 1 in the regressor reflects a change of 10 in the long-run vintage-adjusted citation percentile spread (columns 1 and 5), disruption scaled so a change of 1 in the regressor reflects a change of 10 in the maximum disruption index percentile (columns 2 and 6), pivoting (columns 3 and 7), and standing out (columns 4 and 8). All estimates include an indicator variable for above the 90%ile in novelty as well as an interaction term between this indicator variable and the risk-taking measure. Robust standard errors were used, clustered at the investigator and grant. [†] $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

C. Novelty above the 95%ile

	Base controls				Grant output controls			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Novelty >95%ile	0.0323** (0.0119)	0.0333* (0.0152)	0.1700** (0.0390)	0.2692** (0.0605)	0.0371** (0.0117)	0.0298* (0.0148)	0.1599** (0.0381)	0.2371** (0.0585)
Extreme tail outcomes	-0.0127** (0.0006)				-0.0072** (0.0007)			
Extreme tail outcomes × Novelty >95%ile	0.0028 (0.0025)				0.0014 (0.0025)			
Disruption		-0.0149** (0.0006)				-0.0081** (0.0007)		
Disruption × Novelty >95%ile		0.0022 (0.0024)				0.0019 (0.0024)		
Pivoting			-0.1841** (0.0099)				-0.2340** (0.0099)	
Pivoting × Novelty >95%ile			-0.1221** (0.0471)				-0.1241** (0.0461)	
Standing out				-0.2204** (0.0179)				-0.1815** (0.0177)
Standing out × Novelty >95%ile				-0.2244** (0.0668)				-0.2036** (0.0647)

Note: Estimates stem from a linear probability model where the unit of analysis is the investigator-grant-cycle and the dependent variable is an indicator variable for if the grant cycle was renewed. Columns with base controls are very similar to Table 2, column 2, while those with grant output controls are very similar to Table 2, column 3. Each estimate includes a risk-taking measure as a continuous variable: extreme tail outcomes scaled so a change of 1 in the regressor reflects a change of 10 in the long-run vintage-adjusted citation percentile spread (columns 1 and 5), disruption scaled so a change of 1 in the regressor reflects a change of 10 in the maximum disruption index percentile (columns 2 and 6), pivoting (columns 3 and 7), and standing out (columns 4 and 8). All estimates include an indicator variable for above the 95%ile in novelty as well as an interaction term between this indicator variable and the risk-taking measure. Robust standard errors were used, clustered at the investigator and grant. [†] $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

Table A2. Descriptive statistics: Alternative measures of risk taking and novelty

	Renewed grant cycles					Not renewed grant cycles				
	Mean	Median	Std. Dev.	Min.	Max.	Mean	Median	Std. Dev.	Min.	Max.
Range, median-min citation percentile	32.79	31	18.58	0	97	31.98	30	18.29	0	96
Range, max-median citation percentile	17.19	15	11.68	0	84	19.76	18	12.62	0	92
Range, winsorized max-min citation percentile	39.33	38	18.92	1	99	41.65	41	18.66	1	99
Variance, citation percentile	377.65	311	334.37	0	4,579	424.31	354	369.50	0	4,950
Disruption index percentile, mean	39.71	39	15.43	0	100	43.96	44	16.10	0	100
Disruption index without Nk term percentile, max	60.28	63	22.18	0	94	62.55	66	22.57	0	94
Fraction new MeSH vs. investigator's prior work	0.40	0	0.16	0	1	0.42	0	0.16	0	1
Cosine dissimilarity vs. investigator's prior work	0.54	1	0.20	0	1	0.58	1	0.20	0	1
Fraction unique MeSH vs. other investigators	0.55	1	0.23	0	1	0.50	0	0.22	0	1
Cosine dissimilarity vs. other investigators	0.87	1	0.08	0	1	0.88	1	0.08	0	1
MeSH pair age, mean	25.58	25	7.48	0	63	28.30	28	8.18	1	67
MeSH age, mean	32.73	32	8.37	0	67	35.68	36	9.03	4	64

Note: Outcomes are measured at the investigator-grant-cycle level. Measures are only defined for those cycles with associated publications (see footnote 7 for details).

Table A3. Heterogeneity by degree of pivoting in proposed grant abstract

	Full sample		Greater overlap between proposal and prior work		Less overlap between proposal and prior work	
	Base controls	Grant output controls	Base controls	Grant output controls	Base controls	Grant output controls
<10%ile, fraction new MeSH pairs	0.053** (0.004)	0.059** (0.004)	0.054** (0.006)	0.057** (0.005)	0.063** (0.008)	0.059** (0.008)
>90%ile, fraction new MeSH pairs	-0.024** (0.004)	-0.043** (0.004)	-0.032** (0.010)	-0.062** (0.010)	-0.017** (0.005)	-0.032** (0.005)
Mean of dependent variable	0.455	0.455	0.409	0.409	0.395	0.395
Std. Dev. of dependent variable	0.498	0.498	0.492	0.492	0.489	0.489
Effect bottom %ile group, in s.d. units	0.107	0.119	0.109	0.116	0.129	0.121
Effect top %ile group, in s.d. units	-0.048	-0.086	-0.064	-0.126	-0.035	-0.065
Adjusted R ²	0.4847	0.5039	0.4720	0.4921	0.4734	0.4911
Nb. of investigators	36,539	36,539	21,618	21,618	25,374	25,374
Nb. of investigator-grant-cycles	101,338	101,338	44,229	44,229	43,004	43,004

Note: Estimates stem from a linear probability model where the unit of analysis is the investigator-grant-cycle and the dependent variable is an indicator variable for if the grant cycle was renewed. Each column reflects the results of a separate regression. Columns with base controls are very similar to Table 2, column 2, while those with grant output controls are very similar to Table 2, column 3. Columns 3 and 4 limit the sample to those investigator-grant-cycles with below median fraction new MeSH terms between the grant proposal abstract and an investigator's publications in the preceding 5 years, while Columns 5 and 6 limit the sample to above the median fraction new MeSH. The independent variables are indicator variables for <10%ile and >90%ile for pivoting, with 10-90%ile as the omitted category. Robust standard errors were used, clustered at the investigator and grant. [†] $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

Table A4. Robustness: Investigator name frequency

	Full sample		Below median name frequency		Above median name frequency	
	Base controls	Grant output controls	Base controls	Grant output controls	Base controls	Grant output controls
Extreme tail outcomes						
<10%ile	0.051** (0.004)	0.022** (0.005)	0.051** (0.006)	0.019** (0.007)	0.055** (0.006)	0.025** (0.007)
>90%ile	-0.044** (0.004)	-0.021** (0.004)	-0.044** (0.006)	-0.018** (0.006)	-0.045** (0.006)	-0.022** (0.006)
Disruption						
<10%ile	0.089** (0.012)	0.054** (0.012)	0.109** (0.016)	0.072** (0.016)	0.072** (0.017)	0.037* (0.017)
>90%ile	-0.022** (0.004)	0.008 [†] (0.004)	-0.027** (0.006)	0.004 (0.006)	-0.019** (0.006)	0.011 [†] (0.006)
Pivoting						
<10%ile	0.053** (0.004)	0.059** (0.004)	0.050** (0.006)	0.056** (0.006)	0.057** (0.006)	0.062** (0.006)
>90%ile	-0.024** (0.004)	-0.043** (0.004)	-0.026** (0.006)	-0.046** (0.006)	-0.021** (0.006)	-0.040** (0.006)
Standing out						
<10%ile	0.038** (0.005)	0.025** (0.005)	0.043** (0.007)	0.027** (0.007)	0.035** (0.007)	0.024** (0.007)
>90%ile	-0.023** (0.005)	-0.027** (0.005)	-0.019** (0.007)	-0.023** (0.007)	-0.029** (0.007)	-0.032** (0.007)

Note: Estimates stem from a linear probability model where the unit of analysis is the investigator-grant-cycle and the dependent variable is an indicator variable for if the grant cycle was renewed. Each column and independent variable measure reflect the results of a separate regression. Columns with base controls are very similar to Table 2, column 2, while those with grant output controls are very similar to Table 2, column 3. Columns 3 and 4 limit the sample to those investigator-grant-cycles with below median name frequency, calculated relative to the corpus of PubMed, while Columns 5 and 6 limit the sample to above median name frequency. The independent variables are indicator variables for <10%ile and >90%ile for each measure of risk taking, with 10-90%ile as the omitted category. Robust standard errors were used, clustered at the investigator and grant. [†] $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

Table A5. Robustness: Reliance on NIH funding

	Full sample		Below median fraction pubs. with NIH support		Above median fraction pubs. with NIH support	
	Base controls	Grant output controls	Base controls	Grant output controls	Base controls	Grant output controls
Extreme tail outcomes						
<10%ile	0.051** (0.004)	0.022** (0.005)	0.047** (0.006)	0.021** (0.006)	0.059** (0.007)	0.024** (0.007)
>90%ile	-0.044** (0.004)	-0.021** (0.004)	-0.044** (0.006)	-0.021** (0.006)	-0.045** (0.005)	-0.021** (0.006)
Disruption						
<10%ile	0.089** (0.012)	0.054** (0.012)	0.083** (0.015)	0.053** (0.015)	0.101** (0.019)	0.050** (0.019)
>90%ile	-0.022** (0.004)	0.008† (0.004)	-0.019** (0.006)	0.009 (0.006)	-0.025** (0.006)	0.003 (0.006)
Pivoting						
<10%ile	0.053** (0.004)	0.059** (0.004)	0.047** (0.006)	0.053** (0.006)	0.062** (0.006)	0.065** (0.006)
>90%ile	-0.024** (0.004)	-0.043** (0.004)	-0.022** (0.005)	-0.041** (0.005)	-0.026** (0.006)	-0.046** (0.006)
Standing out						
<10%ile	0.038** (0.005)	0.025** (0.005)	0.042** (0.007)	0.028** (0.007)	0.035** (0.007)	0.022** (0.007)
>90%ile	-0.023** (0.005)	-0.027** (0.005)	-0.021** (0.007)	-0.024** (0.007)	-0.025** (0.006)	-0.029** (0.006)

Note: Estimates stem from a linear probability model where the unit of analysis is the investigator-grant-cycle and the dependent variable is an indicator variable for if the grant cycle was renewed. Each column and independent variable measure reflect the results of a separate regression. Columns with base controls are very similar to Table 2, column 2, while those with grant output controls are very similar to Table 2, column 3. Columns 3 and 4 limit the sample to those investigator-grant-cycles with below median fraction of career publications in PubMed acknowledging any NIH funding, while Columns 5 and 6 limit the sample to those with above median fraction career publications with any NIH funding. The independent variables are indicator variables for <10%ile and >90%ile for each measure of risk taking, with 10-90%ile as the omitted category. Robust standard errors were used, clustered at the investigator and grant. † $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

Table A6. Robustness: Time period

	Full sample		Below median grant evaluation year		Above median grant evaluation year	
	Base controls	Grant output controls	Base controls	Grant output controls	Base controls	Grant output controls
Extreme tail outcomes						
<10%ile	0.051** (0.004)	0.022** (0.005)	0.064** (0.006)	0.034** (0.007)	0.031** (0.006)	0.013* (0.006)
>90%ile	-0.044** (0.004)	-0.021** (0.004)	-0.050** (0.006)	-0.014* (0.006)	-0.033** (0.005)	-0.022** (0.005)
Disruption						
<10%ile	0.089** (0.012)	0.054** (0.012)	0.081** (0.014)	0.054** (0.014)	0.081** (0.018)	0.039* (0.018)
>90%ile	-0.022** (0.004)	0.008† (0.004)	-0.022** (0.006)	0.014* (0.006)	-0.018** (0.005)	0.002 (0.005)
Pivoting						
<10%ile	0.053** (0.004)	0.059** (0.004)	0.061** (0.006)	0.062** (0.006)	0.038** (0.006)	0.046** (0.005)
>90%ile	-0.024** (0.004)	-0.043** (0.004)	-0.023** (0.006)	-0.040** (0.006)	-0.018** (0.005)	-0.035** (0.005)
Standing out						
<10%ile	0.038** (0.005)	0.025** (0.005)	0.023** (0.007)	0.008 (0.007)	0.052** (0.007)	0.043** (0.007)
>90%ile	-0.023** (0.005)	-0.027** (0.005)	-0.006 (0.006)	-0.011† (0.006)	-0.045** (0.007)	-0.048** (0.007)

Note: Estimates stem from a linear probability model where the unit of analysis is the investigator-grant-cycle and the dependent variable is an indicator variable for if the grant cycle was renewed. Each column and independent variable measure reflect the results of a separate regression. Columns with base controls are very similar to Table 2, column 2, while those with grant output controls are very similar to Table 2, column 3. Columns 3 and 4 limit the sample to those investigator-grant-cycles evaluated for renewal in 2003 and earlier, while Columns 5 and 6 limit the sample to 2004 and later. The independent variables are indicator variables for <10%ile and >90%ile for each measure of risk taking, with 10-90%ile as the omitted category. Robust standard errors were used, clustered at the investigator and grant. † $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

Table A7. Robustness: Oster's Delta

	Base controls		Grant output controls	
	Coefficient estimate	Oster's Delta	Coefficient estimate	Oster's Delta
Extreme tail outcomes	-0.0128** (0.0006)	-4.550	-0.0072** (0.0007)	-16.076
Disruption	-0.0151** (0.0006)	-48.926	-0.0081** (0.0007)	2.675
Pivoting	-0.1727** (0.0097)	-23.281	-0.2262** (0.0097)	-9.879
Standing out	-0.2171** (0.0178)	-0.695	-0.1782** (0.0175)	-0.592

Note: Each cell in columns 1 and 3 stems from a separate linear probability model where the unit of analysis is the investigator-grant-cycle and the dependent variable is an indicator for if the grant cycle was renewed. This is very similar to columns 2 and 3 of Table 2A, 2B, 2C and 2D for the extreme tail outcomes (row 1), disruption (row 2), pivoting (row 3) and standing out (row 4), respectively, except a single continuous regressor is used for each measure of risk taking. Robust standard errors were used, clustered at the investigator and grant. Columns 2 and 4 report the corresponding δ parameter from Oster (2019), which can be interpreted as the degree of selection on unobservables relative to observables necessary to explain away the effect of risk taking on grant renewal. We follow Oster's recommendation of setting $R^{\max} = 1.3 \times R^2$. [†] $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.