Where are All the Scientists?
Resources for Studying the Long-Term Careers of STEM Ph.D.s

Stephanie D. Cheng*

April 2020

Abstract

Despite the considerable time and federal funding poured into training scientists, little attention has been given to the role of graduate programs and postdoctoral appointments on future careers – even as STEM trainees spend longer time in these positions. Basic information – such as the number of postdoctoral researchers at each institution – has proven difficult to collect, and the relevant data is spread across various sources. Thus, to assist meta-researchers, this white paper compiles a list of available resources that can be used to study the long-term career outcomes of STEM Ph.Ds. It also identifies shortcomings in current data collection and possibilities for future research avenues.

I would like to thank the National Institute on Aging through grant R24AG058049 to the National Bureau of Economic Research (NBER) for generous financial support of this research; David Cutler, James Banks, and Janet Stein for organizing the NBER-IFS (Institute of Fiscal Studies) International Network on the Value of Medical Research meeting; and participants of the April meeting for their helpful feedback and comments. All errors are my own. The content of this paper is solely the responsibility of the author and does not necessarily represent the official views of the National Institutes of Health or the NBER.

*Harvard University, Department of Economics, stephaniecheng@g.harvard.edu

Contents

1 Introduction

A considerable amount of federal funding and time is spent training the next generation of scientists. The U.S. annually appropriates \$2.8-\$3.4 billion on science, technology, engineering, and mathematics (STEM) education programs: two programs alone – the National Science Foundation (NSF)'s Graduate Research Fellowships and the National Institutes of Health (NIH)'s Ruth L. Kirschstein National Research Service Awards – contribute \$332 million on supporting graduate students and \$473 million on supporting postdoctoral researchers respectively.(Granovskiy 2018) Each year, U.S. universities confer approximately 45,000 STEM doctorate degrees, who spend on average 6.8 years in graduate school.(*Science and Engineering Indicators* 2018) The majority of STEM doctoral recipients then move into postdoctoral appointments, spending on average 1.73 years in these positions.(Cheng n.d.)

At the same time, the STEM fields are known for having a "leaky" pipeline: only 31.37% of biological science Ph.D.s and 22.08% of chemistry Ph.D.s ever move into a tenure-track position.(Cheng n.d.) Approximately 25% of biomedical Ph.Ds. hold non-research positions outside of academia, and nearly 50% of biomedical Ph.Ds. state their occupation is only somewhat or not at all related to their field of training.(Stephan 2013) These "leaks" are especially prevalent among underrepresented populations: while women constitute approximately 45% of postdoctoral fellows in the biomedical sciences, they make up approximately 29% of tenure-track investigators.(Martinez et al. 2007) Underrepresented minorities make up approximately 11% of biomedical postdoctoral fellows but only 6% of tenure-track professors.(Meyer et al. 2018) This homogenous workforce – especially at the higher levels – can have a detrimental impact, as previous research confirms the importance of diversity on scientific innovation.(Gewin 2018)

Given the extensive federal funding and time poured into training scientists, it is important to address what factors contribute to the leaky pipeline. Thus far, little attention has been given to the role of graduate programs and postdoctoral appointments on future careers - despite the lengthening amount of time scientists spend in these positions. Even basic information - such as the number of postdoctoral researchers at each institution - have proven difficult to collect.(*Biomedical Workforce Working Group Report* 2012) This white paper thus has 2 goals: 1) to compile a list of available resources that can be used in studying the long-term career outcomes of STEM Ph.D.s, and 2) to identify gaps in the literature that could be filled with additional data collection. It is organized as follows: Section 2 describes longitudinal surveys, following the scientific workforce over time. Section 3 describes snapshot of scientist careers at one moment in time. Section 4 describes career-related experiments involving STEM doctorates. Section 5 describes databases of scientist records. Finally, Section 6 concludes with future avenues for data collection and research in scientist careers.

2 Longitudinal Surveys

2.1 NSF Survey of Earned Doctorates (SED) & Survey of Doctorate Recipients (SDR)

These two surveys combined provide the most longitudinal, comprehensive description of the U.S. scientific trainee workforce. Beginning in 1957, the Survey of Earned Doctorates (SED; formerly called the Doctoral Records File) is an annual census of all individuals receiving research doctorates from accredited U.S. institutions in that academic year.[1] Administered when a student applies for graduation, the sur-

[1] An earlier version of the DRF contains limited information - sex, institution, field, and year of doctorate - for Ph.Ds. who graduated in 1920-1956.

vey collects information on doctoral recipients' demographics (including date of birth, country of birth, citizenship, race, sex, marital status, parental status), education through the doctorate, and immediate post-graduation plans. More information about the SED, including annual questionnaires, can be found at https://www.nsf.gov/statistics/srvydoctorates/#sd&qs.

From the SED, a nationally representative subset of individuals receiving their first science, engineering, and health research doctorate is selected to be followed in the longitudinal Survey of Doctorate Recipients (SDR). This biennial survey - which can be linked to SED responses - asks individuals for updates on their educational history (including any additional degrees earned or types of training done), employment (including job changes, occupation, tenure-status, salary, compensation, work activities, and satisfaction), and lifestyle (such as marital or parental status changes). Unfortunately, the SDR does not consistently ask about academically-focused job characteristics, such as number of publications, patents, and government support. Individuals are followed until they reach 76 years of age (or are otherwise unable to respond). More information about the SDR, including annual questionnaires, can be found at https://www.nsf.gov/statistics/srvydoctoratework/#sd&qs.

While the SDR was first implemented in 1973, it has undergone several changes over the years; caution should be used to ensure that longitudinal studies across these survey waves are consistent. The 1993 SDR saw a major redesign: the survey layout was reformatted; questions on post-Ph.D. education, current employment, and demographics were reworded and expanded; and the target population was refocused to only include individuals who received U.S. doctorates in science, engineering and health fields.(*Characteristics of Doctoral Scientists and Engineers in the United States: 1993* 1996) Since then, many of the core questions have remained the same, so year-to-year comparisons can be made among the 1993-2017 waves. However, the sample's included individuals may vary from year to year, as substantial changes have been made to the survey's target population. Starting in 2010, the SDR began to survey individuals who have moved abroad in the International SDR (or ISDR) - rather than dropping them from the sample: for the 2010 and 2013 waves, the sample design accounts for individuals residing outside of the U.S. who received doctorate degrees since 2001; starting in 2015, all SED individuals were included in sampling - regardless of academic year of award or post-graduation reesidency. The 2015 wave also saw a major expansion of the SDR sample from approximately 47,000 individuals to 120,000 individuals. To accomplish that increase in sampling, a new sample was selected from the entire SED: the 2015 wave only includes 16,075 individuals from the 2013 SDR; the remainder was newly selected from the 2013-2015 SED.(Foley 2015)

A limited selection of variables are available for public use and can be downloaded from the Scientists and Engineers Statistical Data System at https://ncsesdata.nsf.gov/datadownload/. For access to restricted use microdata, the NSF has a standardized licensed application with instructions available at https://www.nsf.gov/statistics/license/index.cfm. Note that the application now restricts to waves after the 1993 redesign; individuals seeking earlier SDR data may need to separately contact the NSF. Given the extensive data collected and relative ease of obtaining a license, the SED and SDR are popular resources for researchers studying the careers of scientists: Ginther and Kahn 2017 utilize the 1981-2013 waves of the SDR matched to the 1980-2013 SED to examine the impact starting in a postdoctoral position on the employment sector and salaries of biomedical Ph.D.s; they estimate ex-postdocs gave up 17-21% of their present value of income over the first 15 years of their careers relative to Ph.Ds. with no postdoctoral experience. Lan 2012 uses the SED to examine the impact of increased permanent visas through the Chinese Student Protection Act of 1992 on postdoctoral participation; he finds that permanent visa holders are 24% less likely to take postdoctoral positions than temporary visa holders. Kahn and Macgarvie 2018 combine data from the 2010-

2015 waves of the ISDR with country-based limits on EB-2 green cards to estimate the relationship between visa delays and stay rates of international doctorates: each year of visa delay leads to a 2.4 percentage point decline in Chinese graduate stay rates, while Indian graduate students are only affected by very long delays (those facing >5.5 years of delay have a 8.9 percentage point lower stay rate). Agarwal and Ohyama 2013 use the 1995-2006 SDR to fit a life cycle model of human capital investments sorting heterogeneous scientists into different career trajectories; they find evidence of sorting by ability for basic over applied academic research and sorting by nonmonetary returns into academia over industry. Mishagina 2009 examines the occupational choices of science and engineering doctorates - in particular, retention in STEM fields - using the 1973-2001 SDR and 1957-2005 SED; she finds that while 72% of doctorates start their careers in R&D tasks, only 45% were still in R&D 30 years later - with 80% of switchers moving into applied tasks.

2.2 Science & Engineering Ph.D. & Postdoctoral Survey (SEPPS)

Filling in gaps about science trainees' preferences, expectations, and abilities, Roach and Sauermann 2016 administer the extensive Science & Engineering Ph.D. & Postdoctoral Survey (SEPPS) to nearly 6,000 Ph.D. candidates across 39 research-intensive universities and 5 major STEM fields, following them in 2010, 2013, and 2016. SEPPS's longitudinal structure allows them to examine individuals from the early vs. late stages of their Ph.D. (as in Roach and Sauermann 2017); from Ph.D. to postdoc (as in Roach and Sauermann 2016); and a smaller sample from postdoc to postdoc. The survey covers a wide range of measures including career preferences; objective ability (e.g. number of publications, patents, and fellowships); subjective ability (e.g. self-reported research ability relative to peers); expectations about the job market (e.g. percent of field on tenure-track 5 years post-graduation, expected salary); expectations about one's own career (e.g. probability of being on tenure-track within 5 years); and reasons for pursuing postdoctoral positions. Of particular note is how the surveys shed light on non-academic and even non-research careers: Roach and Sauermann 2014 find that over 1/3 of Ph.D. candidates most likely to seek positions in industrial research are not willing to take a lower salary for the opportunity to publish, and Roach and Sauermann 2017 find that 20% of early Ph.D.s are not interested in academic careers - rising to 45% of individuals later in their Ph.D.s. To shed more light on this matter, the surveys also include questions on Ph.D. and postdoc interest in industry careers and entrepreneurship. The authors have generously provided a public-use dataset at http://dx.doi.org/10.7910/DVN/DHSM1F; for further information on the survey, researchers should contact the authors directly.

2.3 National Postdoc Association (NPA) Survey

To better understand the institutional context of postdoctoral researchers, the National Postdoc Association (NPA) survey the resources available to postdocs at each member university. The NPA survey is distributed to postdoctoral offices at NPA's member institutions; 74 institutions completed the 2013 wave, and 102 completed the 2016 wave. The survey asks about institutional and postdoctoral population demographics; structure of the institution's postdoc office; postdoc policies (e.g. term limits, exit survey practices); minimum postdoc stipend policies (in particular, whether institutions adopt the NIH recommended stipend scale - see Subsection 5.5); postdoc benefits (e.g. health insurance, maternity/paternity leave, retirement plans); and professional development/training offerings. The 2016 wave also overlapped with expected changes to the Fair Labor Standards Act (FLSA), which would have increased minimum postdoctoral stipends but was overturned shortly before implementation in December 2016; in response, the NPA sent a follow-up questionnaire to its member institutions to confirm if there were any changes to their responses on postdoctoral compensation. The survey results are detailed in the 2014 and 2017 NPA Institutional Policy Reports; researchers interested in the using the institution-level data should contact the NPA directly.(Ferguson, Huang, et al. 2014; Ferguson, McTighe, et al. 2017) While these surveys do not survey individual postdocs, this institution-level data could be merged with individual-level data to form a more complete picture of their postdoctoral appointments.

3 Snapshot Surveys

3.1 NSF Survey of Graduate Students and Postdoctorates in Science and Engineering (GSS)

While the NSF SED and SDR focuses on individuals who received their doctorates from U.S. universities, which previous research indicates misses a significant proportion of foreign-born STEM doctorates working in the U.S.,[2] the NSF Survey of Graduate Students and Postdoctorates in Science and Engineering (GSS) is an annual count of all research-based graduate students, postdoctoral appointees, and doctorate-level nonfaculty researchers at U.S. universities - regardless of where they received their degrees. The data is publicly available at https://www.nsf.gov/statistics/srvygradpostdoc/pub_data.cfm and can be used to assess general shifts in graduate enrollment and postdoctoral appointments.

The GSS does have several limitations: the survey is limited in scope to each university's tabulations by field of study, U.S. citizenship status, race/ethnicity, gender, part-time or full-time status, and largest mechanism of financial support. It does not include non-academic institutions - such as research centers and federal agencies - and prior to 2017, did not distinguish between Master's and Ph.D. programs in counting graduate students. Because the GSS is distributed to academic institutions and not individuals, it only gives aggregate information and is dependent on the academic institution keeping an accurate count of the number of researchers at their facilities. This is especially problematic in counting the number of postdoctoral appointees, which may be classified under different titles (e.g. "postdoc" vs. "fellow") at different universities; are transient in nature; and - particularly if postdoctoral hiring is handled solely by principal investigators - may not be consistently tracked by universities. In 2010, the GSS was redesigned to improve the accuracy of postdoctoral counts, though there remain concerns that the GSS may still be underestimating the total number of postdoctoral researchers.(Einaudi et al. 2013; Pickett et al. 2017) With these caveats in mind, the data is one of the longest running surveys on the U.S. science trainees and thus gives a good sketch of long-term general trends in the scientific labor force.

3.2 Job Preferences: Stern 2004

Stern 2004 surveys Ph.D. biologists who are (just) completing a job search to determine their preferences for job characteristics - in particular, their willingness to trade off a higher salary for more science-oriented jobs. The survey contains five parts: 1) resume information about the respondent's background and demographics; 2) length and outcome of job search; 3) comparing job offers and an ordinal ranking of offers; 4) cardinal comparison (generally in magnitude and intensity of characteristics) of each individual offer; and 5) ranking of the importance of different job characteristics. The survey was distributed to current postdoctoral researchers

[2]The 2005 Sigma Xi survey (see Subsection 3.3) estimates that 79% of foreign-born postdoctorates working in the U.S. received their doctorates outside of the U.S.(Davis 2005)

whose funding was expiring at four U.S. research institutions; participants in two American Association for the Advancement of Science (AAAS)-sponsored Biology Job Fairs in Cambridge, MA and Palo Alto, CA; and post-Ph.D. biologists with resumes posted to www.biomednet.com. While the overall dataset consists of 107 biologists receiving a total of 223 job offers, the paper focuses on individuals who received multiple research job offers. 66 individuals had multiple job offers; this allows for applicant fixed effects, controlling for heterogeneity such as overall ability or attractiveness to employers. Eliminating non-research jobs - such as management consulting or lab management - gives more similar job comparisons, reducing the sample down to 164 job offers. Because some individuals only completed the ordinal or cardinal comparisons between jobs, the analysis separates into two samples: a cardinal sample of 121 job offers across 52 individuals and an ordinal sample of 134 job offers across 51 individuals. Using this reduced sample, Stern 2004 finds that a 1 standard deviation increase in "science index" - defined as a linear combination of the job's allowance for publishing in external journals, Likert scale rating of incentives to publish in refereed outside journals, and allowance for continuation of current research project as a postdoc - is associated with a more than 6% reduction in predicted wage. He does note that - relative to the variance associated with each of the measured job characteristics - the single-offer averages are not substantially different from the multiple-offer averages. Thus, it may be possible to utilize the remaining single-offer sample in further analysis.

3.3 Postdoc Experience: Sigma Xi, National Postdoc Survey (NPS)

The 2005 Sigma Xi survey - one of the first major U.S. postdoctoral surveys - collects information on 7,600 postdoctoral researchers from 46 institutions, including 18 of the top 20 academic employers and the National Institutes of Health (NIH).(Davis 2005) The questionnaire asks respondents about their demographics (race, ethnicity, citizenship, location obtained doctorate, age, and family structure); postdoctoral satisfaction; salaries and benefits; career expectations; mentorship; and postdoctoral administration. Approximately 1/3 of the survey were considered "core" questions and asked of all respondents; to manage the time needed to fill out the survey, the remaining questions were randomly administered within each institution's participating population. These questions have been made available on the Sigma Xi website at http://postdoc.sigmaxi.org/questions. Unfortunately, due to a system issue, the raw survey data is no longer available to researchers. While the Sigma Xi survey ultimately was a one-shot survey, its questionnaire can provide inspiration for future postdoctoral surveys. Researchers may find it helpful to examine Sigma Xi's extensive list of survey questions in designing their own postdoctoral surveys.

One of the surveys inspired by Sigma Xi, the National Postdoc Survey (NPS) is a postdoctoral survey "designed from a postdoc perspective." Created by postdoctoral researchers primarily associated with the University of Chicago, the 2016 inaugural survey contains responses from 7,603 primarily life science postdoctoral researchers at 351 U.S. institutions; a second wave wrapped up on December 31, 2019. Compared to previously mentioned postdoctoral surveys, the NPS focuses more on asking about the postdoc-PI relationship; availability of professional development programs; finances, benefits, and cost of living; and postdoc satisfaction. It also asks about demographics; grants and publications; job market perceptions and career plans (including back-up plans); and reasons for taking on postdoctoral positions. The results of the 2016 survey are documented in McConnell et al. 2018; they find that formal mentorship training is positively correlated with postdoctoral satisfaction and preference for mentor's career choice. The paper also goes into depth on the protocol and includes the survey instrument in their "Additional Files" section. Summary data for institutions, fields, and regions with more than 50 respondents are available upon request; researchers should contact the study authors for more information.

4 Experiments

4.1 NIH Broadening Experiences in Scientific Training (BEST)

In 2013, the NIH created the Director's Broadening Experiences in Scientific Training (BEST) program. Institutions awarded a 5-year BEST grant implement an experimental training opportunity to prepare biomedical graduate students and postdoctoral researchers for a variety of - particularly non-academic - career options. 10 awards were made in 2013, followed by another 7 in 2014.(*Coalition of Next Generation Life Sciences* n.d.)[3] While each program is individualized to the institution, they primarily used a combination of the following tools: having trainees fill out Individual Development Plan (IDP); offering general skills and professional development workshops (e.g. leadership, communication); holding seminars geared towards specific career paths (e.g. entrepreneurship, pharma); outside mentorship; and short-term shadowing or internship experiences. Additionally, BEST encourages its institutions to track the career outcomes of their biomedical graduate students and postdoctoral researchers over time. (At least two institutions have published the results of such tracking: Wayne State University in Mathur, Cano, et al. 2018 and University of California San Francisco in Silva, Jarlais, et al. 2016.) Program evaluations and related publications are shared on the BEST consortium website at http://www.nihbest.org/publications/. Because most programs are open to all biomedical graduate students and postdoctoral researchers at the university, BEST program evaluations tend to take differences in pre- and post- program surveys or interviews and correlate with demographics (such as gender, race, GRE score, etc.).(Mathur, Chow, et al. 2018; Petrie et al. 2017) There are some possibilities for further rigorous causal estimation: for example, Emory and Georgia Tech's combined program uses a cohort model, which takes in 30 new Ph.D. and postdoctoral scientists a year - leaving the remainder as a possible control group. Programs offering internships also tend to offer differing levels of involvement - from 1-day shadowing to 6-month internships - which may allow for the testing of exposure effects not yet calculated by the current evaluations.(Schnoes et al. 2018; Chatterjee et al. 2019). Future partnership with BEST institutions may result in further understanding of the causal impacts of these career training programs.

4.2 Hypothetical Choices: Ganguli and Gaulé 2018, Janger and Nowotny 2016

Two papers utilize hypothetical choice experiments to measure scientists' willingness to pay for certain job features - in particular, being an academic. In their 2017 survey of 1,605 current chemistry doctoral students, Ganguli and Gaulé 2018 ask respondents to imagine they have multiple job offers and select the percent chance (out of 100) they would accept one offer over the other. To test the respondent's preference for academic positions, the three hypothetical job offers are 1) research scientist at industry firm; 2) postdoctoral researcher at top U.S. university; and 3) teaching-focused assistant professor. To test the respondent's preference for location, they ask respondents to choose between two postdoctoral job offers that differ in either a U.S. university or a foreign university. The authors find that the mean probability of choosing the industry option is approximately 50% for both U.S. and foreign students; of choosing the postdoctoral appointment is 10.2 percentage points higher for foreign students (33.0% vs. 22.8%); and of choosing the teaching assistant professorship is 9.3 percentage points higher for U.S. students (26.2% vs. 16.9%). They also find that on average, foreign students actually have a 12.4 percentage point stronger preference for U.S. postdoctoral positions than foreign positions (60.5% vs. 48.1%). These trends hold even when controlling for graduate

[3]The 17 BEST sites and their program summaries are available here: http://www.nihbest.org/about/17-research-sites/

school, gender, marital status, enrollment year, and field of study.

Similarly, Janger and Nowotny 2016 utilize the hypothetical choice methodology in a large-scale survey of more than 10,000 European researchers across different career stages. Part of the EU-funded "Mobility of Researchers 2" (MORE2) project, the survey asks 3,790 early-stage researchers and 6,425 later-stage, independent researchers for their choice between 3 randomly allocated, academic jobs.(*Support for continued data collection and analysis concerning mobility patterns and career paths of researchers* 2013) The job choices vary in salaries and benefits; country quality of life relative to the country the respondent currently working; and job characteristics (e.g. time for own research, funding, and opportunities for career advancement). The authors find that at average wages, a $1,000 wage increase raises the probability of choosing a job offer by approximately 0.8 percentage points for early-stage researchers and 0.9 percentage points for later-stage researchers. Using the coefficients of a conditional logit regression, the authors calculate the willingness to pay for various job features: in particular, early-stage researchers are willing to pay $2,100 for each additional contract year; $18,659 for tenure possible contigent on performance and job availability; and $21,026 for tenure contingent purely on research performance.

5 Databases

5.1 IRIS UMETRICS

Hosted by the University of Michigan's Institute for Research on Innovation & Science (IRIS), the "Universities: Measuring the Impacts of Research on Innovation, Competitiveness, and Science" (UMETRICS) project collects administrative data from over 30 member universities[4] to examine the social and economic impact of academic research. The core files contain university-sponsored award and grant level data on project expenditures; direct employee wages; vendor purchases; and subaward transactions. Employee data can be linked to ProQuest dissertation data (see Subsection 5.4), publications, patents, NSF SED data (see Subsection 2.1), and Census earnings data. Awards can be linked to their grants' original application data through partnerships with the NIH, NSF, and the US Department of Agriculture. Since its inception in 2013, approximately 100 researchers have accessed the UMETRICS data; among other projects, they have analyzed the earnings outcomes of Ph.D. recipients (Zolas et al. 2015); the relationship between geographic proximity of vendors and university research expenditures (Goldschlag et al. 2019); and the impact of declining federal R&D funding on the organization of research groups (Funk et al. 2019). Researchers interested in using the UMETRICS data can apply through their online form at https://iris.isr.umich.edu/research-data/access/. Individuals affiliated with IRIS member institutions can access the data for free, while non-IRIS affiliated individuals are charged a non-refundable seat fee of $1,250 ($625 for students). Approved projects can then access deidentified data through a secure virtual data enclave.

5.2 Coalition of Next Generation Life Sciences (CNGLS)

In December 2017, the Coalition of Next Generation Life Sciences (CNGLS) was founded with the goal of providing career transparency for life science trainees. Over 50 member institutions have pledged to publicly release data on the career outcomes of their life science Ph.Ds. and postdoctoral researchers - including the admissions and matriculation of Ph.D. students; median time-to-degree and completion data for Ph.D. programs; Ph.D. and postdoctoral demographics - particularly by gender, underrepresented minority status,

[4] A full list of IRIS members can be found at https://iris.isr.umich.edu/iris-members-map/.

and citizenship status; median time in postdoctoral positions at the institution; and Ph.D. and postdoctoral alumni careers. CNGLS provides member institutions reporting guidelines, which allows for cross-institution comparisons. Of particular note is work done by Silva, Mejía, et al. 2019 at the University of California San Francisco (UCSF): meta-researchers may find their paper a helpful blueprint for how to collect and categorize the career outcomes of Ph.D. and postdoctoral career outcomes. Member institutions host their data on their own websites, which is linked on the CNGLS website at http://nglscoalition.org/coalition-data/. Unfortunately, universities do not provide the raw data counts; instead, most provide data visualizations through Tableau graphs - oftentimes across multiple webpages. Thus, meta-researchers hoping to use data from CNGLS institutions may need to scrape the information off each individual institution website.

5.3 Grants: Research Portfolio Online Reporting Tools Expenditures & Results (RePORTER)

As part of the federal government's goals for public transparency and accountability, information on research projects funded by select agencies[5] can be accessed through their online repository, the Research Portfolio Online Reporting Tools Expenditures and Results (RePORTER). This system gives yearly funding success rates - defined as the percentage of reviewed grant applications that receive funding - and allows the general public to query for the projects, publications, patents, and clinical studies tied to each grant. Meta-researchers can also take advantage of downloading bulk RePORTER data through their ExPORTER system, which conveniently packages information on all funded projects in each fiscal year since 1985.[6] For each project, ExPORTER collects information on the principal investigators' names; project title and abstract; grant type; administering institute or center; budget start and end dates; grantee organization; total cost (as well as divided into direct and indirect costs). It links to MEDLINE and PubMed publication data (see Subsection 5.4 for more detail; RePORTER data includes author list, journal information, and publication date); federal patent data (patent ID and title); and clinical studies (title, ClinicalTrials.gov ID, and current stage). Publication data is refreshed every year, while patents and clinical studies data are refreshed every week.

Thus far, the RePORTER data has been extensively used to examine the relationship between research funding and outputs - for example, how targeted grant opportunities can shift scientists' research direction (Myers 2019); how interruptions in grant funding affect scientists' research activity (Tham 2019); and the direct and indirect channels through which federal funding affects patenting (Li et al. 2017). In order to study the impact of federal funding on the careers of established scientists, RePORTER data at the principal investigator level could feasibly be linked to CVs and other career outcomes. For example, Azoulay et al. 2017 pull together an extensive number of data sources - including RePORTER data - to follow 10,051 elite life scientists over time; they find that scientists who've recently received NIH funding are less likely to move, which they attribute to the high transaction costs of transferring funds between institutions. To study the impact on science trainees would be more time-intensive, as only the principal investigators' are listed on RePORTER's project information. However, it may be possible to do so by linking principal investigators to their lab employees at the time of funding or by identifying trainees from linked publication data.

For projects outside of U.S.-funded life sciences, a Federal RePORTER and a World RePORT system

[5]This is primarily the NIH, but the system also includes grants funded by the Administration for Children and Families (ACF), Agency for Healthcare Research and Quality (AHRQ), Centers for Disease Control and Prevention (CDC), Health Resources and Services Administration (HRSA), U.S. Food and Drug Administration (FDA), and Veterans Affairs (VA).

[6]Additionally, ExPORTER's predecessor, CRISP, contains project data from FY1970-2009. However, CRISP data is not linked to publications, patents, or clinical studies data.

have been established. Since fiscal year 2000, the Federal RePORTER annually collects funding data from the Department of Defense (DOD), Department of Education (DOE), Environmental Portection Agency (EPA), Department of Health and Human Services (HHS), National Aeronautics and Space Administration (NASA), and National Science Foundation (NSF). Publications are linked to project data from the EPA; NSF; and select HHS, DOD, and USDA departments. Thus far, Federal RePORTER data is not linked to patents data. Since fiscal year 2012, the World RePORT highlights biomedical research investments from some of the world's largest funding organizations: it currently includes the Bill & Melinda Gates Foundation, the Canadian Institutes of Health Research, European Commission, European & Developing Countries Clinical Trials Partnership, Medical Research Council, Institut Pasteur, Swedish International Development Cooperation Agency, Swedish Research Council, and Wellcome Trust. However, the World RePORT currently does not link projects funded by these organizations to their publications or patents. While these two online repositories are limited compared to the NIH RePORTER, future expansions may allow for more extensive study - especially of non-biomedical and non-U.S. research.

5.4 Publications & Citations: ProQuest, MEDLINE/PubMed, ORCID, Scopus, Web of Science

Several online databases provide information on scientific publications and their citation history. They are generally used to search a scientist's publication record: one queries the online database with the scientist's name - perhaps joined with their affiliation and field of study to reduce mismatches for common names - and is returned a list of publication names, abstracts, journal information, links to the full text, and yearly citations. Some databases may also allow searches for a scientist's other work, such as clinical trials, conference proceedings, and patents.

Some of the most commonly used databases include ProQuest, MEDLINE/PubMed, ORCID, Scopus, and Web of Science - though several others exist as well. The ProQuest Dissertations and Theses database is the largest repository of primarily U.S. graduate dissertations and theses, containing over 4 million theses from over 3,000 universities. Each year, more than 130,000 works are added to the database. In addition to the full-texts (primarily for theses from 1997 onward), ProQuest includes metadata such as the author name, advisor name, committee members, department, university, publication year, and degree date. MEDLINE is the National Library of Medicine's journal citation database, containing over 26 million references to biomedical publications from more than 5,200 journals since 1946; it is primarily accessed through the freely available PubMed, which includes additional citation databases for more than 30 million references. For biomedical meta-research, the premiere database is MEDLINE/PubMed; for example, it is the source from which RePORTER (see Subsection 5.3) links funding to publication data. For fields beyond biomedical research, Scopus and Web of Science are two subscription-based services that cover a wide variety of academic fields. Scopus contains approximately 1.4 billion citations from more than 24,600 journals and 5,000 publishers since 1970; it also automatically constructs author profiles - of which it has approximately 16 million. Web of Science contains approximately 1.7 billion citations from over 21,100 high-impact journals since 1900. In comparison to Scopus, Web of Science focuses on "high influence" publications and covers fewer non-U.S. and interdisciplinary research. Most researchers will obtain access through their academic institutions, which typically subscribes to one of the two.

One drawback of these online databases is the possibility of mismatch between authors and publications due to common names. To correct for this, ORCID is an online database that provides scientists with a persistent digital identifier that they can connect to their affiliations, grants, publications, etc. Because researchers must register for an identifier and link to their work themselves - rather than being automatically generated - ORCID is smaller than the other databases mentioned. Its advantage as its network grows is that it provides a more complete publication work with less possibility of mismatch.

On their own, publication and citation data are limited in their research usability. They are generally combined with additional scientist information as proxies for scientific ability or productivity. For example, Ross et al. 2019 measure how research quality - as proxied by the number of citations and publications - declines over a scientist's career. Buffington et al. 2016 identify Ph.D. recipients in the UMETRICS data (see Subsection 5.1) using ProQuest's Dissertation and Thesis Database, then matches onto career placement and earnings data from the Census; they find that - once broad dissertation topic and funding source are controlled for - there is no gender difference in the likelihood of working in lucrative sectors and that the estimated wage gap drops by about two-thirds to 11 percent. To examine Ph.D.s' preferences over employment outcomes through revealed preference, Conti and Visentin 2015 utilize Scopus publication data to determine the within-field research ranking of universities; the R&D intensiveness of companies; and the research quality of trainees and their supervisors. Balsmeier and Pellens 2014 use cumulative number of publications in the Web of Science database as a proxy for scientist productivity, finding that having an additional publication decreases a scientist's propensity to leave academe by 6%.

5.5 Stipends: NIH Guidelines, Future of Research, PhDStipends.com, & PostdocSalaries.com

Meta-researchers may be interested in looking at Ph.D. and postdoctoral stipends, especially as science advocacy groups argue low stipends may disincentivize talented individuals from remaining in STEM fields. For the biomedical field, a good starting point is the NIH Ruth L. Kirschstein National Research Service Award (NRSA) postdoctoral stipend guidelines. These give the NIH recommended stipend amount for a postdoctoral researcher with a certain number of years of experience at the institution. The NIH has published their historical stipend guidelines on their website and releases an announcement of annual stipend amounts each year.(*Kirschstein-NRSA Stipend History* 2016; *Ruth L. Kirschstein National Research Service Award (NRSA) Stipends, Tuition/Fees and Other Budgetary Levels Effective for Fiscal Year 2020* 2020). Deemed the "gold standard" for minimum stipend amounts, many institutions peg their postdoctoral stipends to the NIH guidelines - even for postdoctoral researchers not funded by the NIH.(Ferguson, McTighe, et al. 2017) In their Institutional Policy Reports (see Subsection 2.3), the NPA estimates that 52% of their member institutions set minimum postdoctoral stipends to the NIH NRSA amount in 2013; even with a large increase to the NIH NRSA stipend guideline in 2016, 61% of their member institutions continued to peg their postdoctoral stipends to the NIH NRSA amount.(Ferguson, Huang, et al. 2014,Ferguson, McTighe, et al. 2017).

Several grassroots advocacy groups have also collected information on science trainee stipends to improve transparency. In 2016, Boston-based Future of Research submitted Freedom of Information Act (FOIA) requests to U.S. public institutions with at least 300 postdoctoral researchers, obtaining the salaries and job titles for over 13,000 postdoctoral researchers at 52 public U.S. institutions; additionally, 1 private university - Boston University - contributed stipend data. They find that 22.7% of postdoctoral researchers had salaries within $25 of the NIH NRSA minimum stipend of $47,484, with 61.0% of postdoctoral researchers having stipends between $40,000-$49,999.99.(Athanasiadou. et al. 2017) The group also publicly provides the de-identified, individual-level data on job title, university, and stipend on their website at https://www.futureofresearch.org/investigating-postdoc-salaries/.

However, because their dataset is mostly limited to postdoctoral researchers at public institutions, Future of Research encourages Ph.D. and postdoctoral researchers to anonymously submit their historical and current stipend information to PhDStipends.com and PostdocSalaries.com, which both publicly display the results of submitted stipend information on their website. PhDStipends.com has over 8,000 submissions that give university; department; overall pay; living wage ratio (measured using the Poverty in America Living Wage Calculator for a single person with no dependents); academic year; program year; and any additional comments. PostdocSalaries.com has over 1,500 submissions that give institution; department; title; salary; living wage ratio; benefits; whether this is a negotiated offer; academic year; years since Ph.D.; whether the institution continues the postdoctoral researcher an employee; and any additional comments.

5.6 Academic Family Tree

Several efforts have been made to link STEM trainees to their mentors, creating an academic genealogy of researchers. The largest - the Academic Family Tree - started with neuroscience in January 2005 and has since expanded to approximately 748,100 people across over 60 fields of study.[7] This open-source database links scientists from mentor to mentee in an intuitive "family tree" structure, with start and end dates of training for each mentee. It also connects to their publication history through the PubMed database and to their NIH and NSF funding data through the Federal RePORTER (Star Metrics) system.[8] Using the Academic Family Tree, Liénard et al. 2018 examine the impact of graduate and postdoctoral mentorship by examining 18,856 "triples" of researchers - consisting of a trainee, a graduate mentor, and a postdoctoral mentor; they find that the postdoctoral mentor has a larger influence than the graduate mentor on a trainee's odds of continuing in academia and own training of new scientists.

There are a few caveats with using the Academic Family Tree for further research. Because the Academic Family Tree is open-source, it depends on user inputs to identify mentor-mentee relationships and thus is not a universal representation of academic relationships. The trees also tend to focus on academic research relationships, so STEM trainees who leave academia or are no longer doing academic research are less likely to be in this database. As the number of Academic Family Tree contributors grows, it may be able to provide a more complete picture of mentor-mentee relationships.

5.7 Diverse Scientists: Request a Woman in STEMM, CAISE

For researchers studying diversity in STEM, there are several databases listing female and minority scientists. Originally created to encourage more diversity in seminar speakers and journalism quotes, meta-researchers could potentially draw from these databases to create scientist samples. The Database of Databases of Diverse Speakers in STEM acts a starting point, compiling a list of databases that collect information on underrepresented groups in STEM. Each database may focus on a different subset of research fields and included groups. For example, the Request a Woman in STEMM (formerly "Request a Woman Scientist") database consists of over 7,500 women from 174 scientific disciplines and 133 countries.(McCullagh et al. 2019) Women in science, technology, engineering, mathematics, and medicine (STEMM) can provide their contact information, career stage, degree, scientific discipline, geographic location, self-identifying dimensions of representation, and professional availability (i.e. willingness to be a seminar speaker, speak to a journalist,

[7] They have linked to other Academic Family Tree projects at https://neurotree.org/neurotree/faq.php, such as the Mathematics Genealogy Project, the Family Tree of Trade Economists, and Brown University's planetary geology family tree.

[8] As PubMed and RePORTER primarily focuses on biomedical fields, these scientists are more likely to list publication and funding data on the Academic Family Tree website.

etc.). Similarly, the Counting All for Inclusion in STEM Equity (CAISE) database collects information on historically marginalized individuals (HMI) with terminal STEM degrees who are currently conducting research at an academic institution. This includes their contact information, HMI identifiers, field of study, a link to their professional website, and professional availability. Given that the nature of these databases to bring awareness to and more easily contact underrepresented groups in STEM, the listed individuals may be willing to contribute to surveys - particularly on diversity in STEM - or can provide a starting sample of scientists to merge onto other datasets, such as their publications, patents, and career paths.

5.8 Professional Associations

In addition to the data sources that have already been mentioned in this white paper, consider partnerships with science advocacy groups, professional associations, and honor societies to generate new data on the careers of scientists. Among many others, this includes the American Association of Arts and Sciences (AAAS), American Society for Biochemistry and Molecular Biology (ASBMB), American Chemical Society (ACS), engineering honor society Tau Beta Pi (TBP), and American Physics Society (APS). Not only can they provide first-hand knowledge about the types of careers pursued in their field, these organizations may have existing career programs that could be leveraged in meta-research experiments. For example, APS offers an industry mentorship program, matching physics graduate students and postdoctoral researchers with physicists who have experience working in industry. Professional associations can also assist in constructing a large sample of STEM trainees. As they maintain contact information for working professionals in their field, they can help meta-researchers distribute surveys or administer other data collection processes to their membership. For example, Sigma Xi and the National Postdoc Survey - mentioned in Subsection 3.3 - have distributed their survey through the National Postdoctoral Association (see Subection 2.3) and by contacting individual institutions' postdoctoral offices. Especially if meta-researchers are interested in a particular field, professional associations are a good starting point to examine what career projects have already been implemented and to contact a large group of professionals in the field.

6 Future Avenues

This white paper has outlined existing sources for studying the long-term career paths of scientists, though each come with their limitations. The majority of surveys focus on individuals in the biomedical sciences who received their Ph.Ds. from U.S. institutions; much is left to be learned about the estimated 71% of STEM Ph.Ds. in other fields and 47% of STEM postdoctoral researchers working in the U.S. who received their degrees abroad.(Cheng n.d., Davis 2005) Several databases can indirectly shed light on the career paths of scientists, piecing together data from federal grants (see Subsection 5.3), publications (see Subection 5.4), mentorship (see Subection 5.6). For a subset of schools, organizations have already begun the work of merging these various databases together; in particular, UMETRICS (see Subsection 5.1) is working on combining lab-level employee data with grants, publications, patents, and Census earnings data. However, more work is needed to construct a nationally representative sample of STEM trainees with complete career paths, ability proxies, and preference measures. Future partnerships with science advocacy and professional associations may assist with constructing such a data sample. While such an endeavor will take a considerable amount of work, it would open many more opportunities to study the long-term careers of scientists.

7 Works Cited

References

Agarwal, R., & Ohyama, A. (2013). Industry or academia, basic or applied? career choices and earnings trajectories of scientists. *Management Science*, *59*(4), 950–970.

Athanasiadou., R., Bankston, A., Carlisle, M., Niziolek, C., & McDowell, G. (2017). Assessing the landscape of u.s. postdoctoral salaries. *bioRxiv*.

Azoulay, P., Ganguli, I., & Zivin, J. G. (2017). The mobility of elite life scientists: Professional and personal determinants. *Research Policy*, *46*(3), 573–590.

Balsmeier, B., & Pellens, M. (2014). Who makes, who breaks: Which scientists stay in academe? *Economics Letters*, *122*(2), 229–232.

Biomedical workforce working group report. (2012). National Institutes of Health.

Buffington, C., Cerf, B., Jones, C., & Weinberg, B. A. (2016). Stem training and early career outcomes of female and male graduate students: Evidence from umetrics data linked to the 2010 census. *American Economic Review: Papers and Proceedings*, *106*(5), 333–338.

Characteristics of doctoral scientists and engineers in the united states: 1993. (1996)(tech. rep. No. NSF 96-302). National Science Foundation. Arlington, VA.

Chatterjee, D., Ford, J. K., Rojewski, J., & Watts, S. W. (2019). Exploring the impact of formal internships on biomedical graduate and postgraduate careers: An interview study. *CBE - Life Sciences Education*, *18*(2).

Cheng, S. D. (n.d.). *The career paths of scientists*. Ongoing work using NSF Survey of Doctoral Recipients, 1993-2013 datasets.

Coalition of Next Generation Life Sciences. (n.d.). Retrieved from https://nglscoalition.org

Conti, A., & Visentin, F. (2015). A revealed preference analysis of phd students' choices over employment outcomes. *Research Policy*, *44*(10), 1931–1947.

Davis, G. (2005). Doctors without orders. *American Scientist*, *93*(3, supplement).

Einaudi, P., Heuer, R., & Green, P. (2013). *Counts of postdoctoral appointees in science, engineering, and health rise with reporting improvements*. National Center for Science and Engineering Statistics.

Ferguson, K., Huang, B., Beckman, L., & Sinche, M. (2014). *Supporting and developing postdoctoral scholars: National postdoctoral association institutional policy report 2014*. National Postdoctoral Association.

Ferguson, K., McTighe, M., Amlani, B., & Costello, T. (2017). *Supporting the needs of postdocs: 2017 national postdoctoral association institutional policy report*. National Postdoctoral Association.

Foley, D. J. (2015). *Survey of doctorate recipients, 2015 (technical notes)*. National Center for Science and Engineering Statistics. Retrieved from https://ncsesdata.nsf.gov/doctoratework/2015/sdr_2015_tech_notes.pdf

Funk, R., Glennon, B., Lane, J., Murciano-Goroff, R., & Ross, M. (2019). Money for something: Braided funding and the structure and output of research groups. *IZA Discussion. Paper*, (12762).

Ganguli, I., & Gaulé, P. (2018). Will the u.s. keep the best and the brightest (as post-docs)? career and location preferences of foreign stem phds. *NBER Working Paper*, (24838).

Gewin, V. (2018). Why diversity helps to produce stronger research. Retrieved from https://www.nature.com/articles/d41586-018-07415-9

Ginther, D. K., & Kahn, S. (2017). The impact of postdoctoral training on early careers in biomedicine. *Nature Biotechnology*, *35*(1), 90–94.

Goldschlag, N., Lane, J., Weinberg, B., & Zolas, N. (2019). Proximity and economic activity: An analysis. of vendor-university transactions. *Journal of Regional Science*, *59*(1), 163–182.

Granovskiy, B. (2018). Science, technology, engineering, and mathematics (STEM) education: An overview. *Congressional Research Service*.

Janger, J., & Nowotny, K. (2016). Job choice in academia. *Research Policy*, *45*(8), 1672–1683.

Kahn, S., & Macgarvie, M. (2018). The impact of permanent residency delays for stem phds: Who leaves and why. *National Bureau of Economic Research*, (25175).

Kirschstein-nrsa stipend history. (2016). National Institutes of Health.

Lan, X. (2012). Permanent visas and temporary jobs: Evidence from postdoctoral participation of foreign phds in the united states. *Journal of Policy Analysis and Management*, *31*(3), 623–640.

Li, D., Azoulay, P., & Sampat, B. N. (2017). The applied value of public investments in biomedical research. *Science*, *356*(6333), 78–81.

Liénard, J. F., Achakulvisut, T., Acuna, D. E., & David, S. V. (2018). Intellectual synthesis in mentorship determines success in academic careers. *Nature Communications*, *9*(4840).

Martinez, E. D., Botos, J., Dohoney, K. M., Geiman, T. M., Kolla, S. S., Olivera, A., ... Cohen-Fix, O. (2007). Falling off the academic bandwagon. women are more likely to quit at the postdoc to principal investigator transition. *EMBO Reports*, *8*(11), 977–981.

Mathur, A., Cano, A., Kohl, M., Muthunayake, N. S., Vaidyanathan, P., Wood, M. E., & Ziyad, M. (2018). Visualization of gender, race, citizenship and academic performance in association with career outcomes of 15-year biomedical doctoral alumni at a public research university. *PLoS ONE*, *13*(5), e0197473.

Mathur, A., Chow, C. S., Feig, A. L., Kenaga, H., Moldenhauer, J. A., Muthunayake, N. S., ... Straub, V. (2018). Exposure to multiple career pathways by biomedical doctoral students at a public research university. *PLoS ONE*, *13*(6), e0199720.

McConnell, S. C., Westerman, E. L., Pierre, J. F., Heckler, E. J., & Schwartz, N. B. (2018). Research: United states national postdoc survey results and the interaction of gender, career choice, and mentor impact. *eLife*, *7:e40189*. Retrieved from https://elifesciences.org/articles/40189

McCullagh, E. A., Nowak, K., Pogoriler, A., Metcalf, J. L., Zaringhalam, M., & Zelikova, T. J. (2019). Request a woman scientist: A database for diversifying the public face of science. *PLoS Biology*, *17*(4), e3000212.

Meyer, L. C., Brown, A. M., Moneta-Koehler, L., & Chalkley, R. (2018). Survey of checkpoints along the pathway to diverse biomedical research faculty. *PLoS ONE*, *13*(1), e0190606.

Mishagina, N. (2009). *Labor market behavior of sciences and engineering doctorates: Three essays* (Doctoral dissertation, Queen's University).

Myers, K. (2019). The elasticity of science. *SSRN*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3176991

Petrie, K. A., Carnahan, R. H., Brown, A. M., & Gould, K. L. (2017). Providing experiential business and management training for biomedical research trainees. *CBE Life Sci Educ*, *16*(3), ar51.

Pickett, C. L., Bankston, A., & McDowell, G. S. (2017). The gss is an unreliable indicator of biological sciences postdoc population trends. *BioRxiv*. Retrieved from https://www.biorxiv.org/content/10.1101/171314v3

Roach, M., & Sauermann, H. (2014). Not all scientists pay to be scientists: Phds' preferences for publishing in industrial employment. *Research Policy*, *43*(1), 32–47.

Roach, M., & Sauermann, H. (2016). Why pursue the postdoc path? *Science*, *352*(6286), 663–664.

Roach, M., & Sauermann, H. (2017). *The declining interest in an academic career.*

Ross, M. B., Yu, H., Marschke, G., Staudt, J., & Weinberg, B. A. (2019). *Publish or perish: Selective attrition as an unifying explanation for lifecycle patterns in innovation.* NBER Productivity Seminar.

Ruth l. kirschstein national research service award (nrsa) stipends, tuition/fees and other budgetary levels effective for fiscal year 2020. (2020). National Institutes of Health. Retrieved from https://grants.nih.gov/grants/guide/notice-files/NOT-OD-20-070.html

Schnoes, A. M., Caliendo, A., Morand, J., Dilinger, T., Naffziger-Hirsch, M., Moses, B., . . . O'Brien, T. C. (2018). Internship experiences contribute to confident career decision making for doctoral students in the life sciences. *CBE - Life Sciences Education*, *17*(1).

Science and Engineering Indicators. (2018). National Science Board, National Science Foundation. Retrieved from https://www.nsf.gov/statistics/indicators/

Silva, E. A., Jarlais, C. D., Lindstaedt, B., Rotman, E., & Watkins, E. S. (2016). Tracking career outcomes for postdoctoral scholar: A call to action. *PLoS ONE*, *15*(2), e1002600.

Silva, E. A., Mejía, A. B., & Watkins, E. S. (2019). Where do our graduates go? a tool kit for tracking career outcomes of biomedical phd students and postdoctoral scholars. *CBE - Life Sciences Education*, *18*(3), 1–6.

Stephan, P. (2013). How to exploit postdocs. *BioScience*, *6*(4), 245–246.

Stern, S. (2004). Do scientists pay to be scientists? *Management Science*, *50*(6), 835–853.

Support for continued data collection and analysis concerning mobility patterns and career paths of researchers. (2013). IDEA Consult. Retrieved from https://cdn2.euraxess.org/sites/default/files/policy_library/report_on_survey_of_researchers_outside_eu.pdf

Tham, W. Y. (2019). *Science, interrupted: Budget uncertainty and its effects on innovation* (Doctoral dissertation, Ohio State University).

Zolas, N., Goldschlag, N., Jarmin, R., Stephan, P., Owen-Smith, J., Rosen, R. F., . . . Lane, J. I. (2015). Wrapping it up in a person: Examining employment and earnings outcomes for phd recipients. *Science*, *350*(6266), 1367–1371.