Early Stage Investigator Policy Evaluation: The Statistical Necessities

Rachael Walsh, PhD, Robert F. Moore, Jamie Mihoko Doyle, PhD, and Katrina Pearson
National Institutes of Health, Office of Extramural Research, 6705 Rockledge Drive, Bethesda, MD 20817

Abstract
To assist new scientists in the transition to independent research careers, the National Institutes of Health (NIH) implemented an Early Stage Investigator (ESI) policy beginning with applications submitted in 2009. During the review process, the ESI designation segregates applications submitted by investigators within 10 years of completing their terminal degree or medical residency from applications submitted by more experienced investigators. Institutes/Centers can then give special consideration to ESI applications when making funding decisions. One goal of this policy is to increase the probability of newly emergent investigators receiving research support. Using direct matching algorithms to generate comparable groups pre- and post-policy implementation, generalized linear models were used to evaluate the ESI policy, comparing the probability of funding for ESI flagged applications from 2011 to 2015 to applications from 2004 to 2008 with similar characteristics. This paper addresses the statistical necessities of public policy evaluation, finding that the ESI policy stabilized the proportion of NIH funded newly emergent investigators. In the absence of the ESI policy, 54 percent of newly emergent investigators would not have received funding.

Keywords: early stage investigator, NIH grant funding, policy evaluation

1. Introduction

Receiving independent research support continues to be an important milestone marking the transition from a newly emergent investigator to an established investigator at many biomedical research institutions in the United States (National Research Council, 2005). However, current trends in biomedical research funding have created a hypercompetitive environment (Cook, Grang, and Eyre-Walker, 2015) and young investigators are attaining research independence later in their careers (Basken and Voosen, 2014; Rockey, 2012). For instance, studies have shown that the average age to first major research grant from the National Institutes of Health (NIH) has been increasing from 36 in 1980 to 42 in 2008 (Matthews, Calhoun, Lo, and Ho, 2011) and reached 45 in 2016 (NIH, 2017). Youth and diversity in the biomedical workforce are linked to major scientific breakthroughs, revolutionizing medicine and the healthcare of the populace (Jones, Reedy, and Weinberg, 2014), with Nobel Laureates in medicine conducting prize-winning work by age 45 (Redelmeier and Naylor, 2016). Reasons for the trend of an aging biomedical research workforce include an oversupply of young scientists relative to the number of open faculty positions (Alberts, Kirschner, Tilghman, and Varmus, 2014; Clauset, Larremore, and Sinatra, 2017), and the presence of more experienced, prolific cohorts of established investigators who disproportionately receive NIH research awards (Levitt and Levitt, 2017). Taken together, these two factors can create serious,

Disclaimer: The views expressed in this paper are those of the authors and do not necessarily represent those of the National Institutes of Health or the United States Department of Health and Human Services.

long term consequences to the biomedical research workforce by forcing young scientists to seek out career opportunities outside of academic research (Daniels, 2017) and limiting support available to scientists at the most creative stage of their careers (Jones, Reedy, and Weinberg, 2014).

As one of the largest sources of financial support for biomedical research in the world (Viergever and Hendriks, 2015), the NIH implemented policies aimed at sustaining a more balanced biomedical research workforce and to lower the increasing age to first major research grant (Heggeness, Carter-Johnson, Schaffer, and Rockey, 2016; Levitt and Levitt, 2017). In 2009, the NIH implemented the Early Stage Investigator (ESI) policy. The purpose of the policy was to, "counter advantages enjoyed by well-established investigators and to encourage early transition to independence." (https://grants.nih.gov/policy/new_investigators/index.htm). To accomplish this, the NIH ESI policy requires ESI-eligible applications to be segregated during review and reviewers are instructed to score the application based on the merits and ideas within the application, and not necessarily focus on the writing, preliminary data, and career stage of the investigator. To qualify for ESI status, all program directors/principal investigators on an application must not have prior substantial NIH independent research awards *and* be within 10 years of his/her terminal degree or end of medical residency.

Despite the implementation of the ESI policy, recent research found older, more experienced NIH awardees are still more likely to have more than one award resulting in enhanced survival benefits within the research project grant (RPG) funding system (Charette et al, 2016). Funding disparities are two-fold – experienced investigators are more likely to have applications funded than ESIs *and* the direct award dollar amount per investigator disproportionately favors experienced investigators (Charette et al, 2016). Additionally, the aging baby boom cohort of scientists in conjunction with the decline in the retirement rate and elimination of mandatory retirement in universities have resulted in a rapidly aging scientific workforce (Blau and Weinberg, 2017). The ESI policy was intended to diminish the advantage of experience and make it easier for newly emergent investigators to transition to independence.

While existing research and published data have suggested that the ESI policy has not reversed trends in obtaining NIH research funding, studies to date have not formally examined the effectiveness of the policy. For instance, the existing body of literature rely on either descriptive statistics (Dorsey and Wallen, 2016; Moore, 2017) or were limited to data from one IC (Berg, 2010; Boyington, Antman, Patel, and Lauer, 2016; NIDDK, 2017). Descriptive statistics are necessary for exploratory data analysis and the first step in any formal analysis; however, descriptive statistics cannot infer causality or evaluate policy. In addition, restricting data to just one IC limits the generalizability of findings. That is, the policy's overall effectiveness cannot be ascertained.

The purpose of this research was to use a quasi-experimental design to infer causality with respect to the ESI policy. More specifically, this research asked the following:

- Can a quasi-experimental design be applied to existing data to infer ESI policy effectiveness?
- If so, are newly emergent investigators more likely to receive funding for applications submitted post-policy implementation when compared to similar applications submitted pre-policy implementation?

The first stage of this research focused on matching post-policy applications to pre-policy applications then evaluated the quality of the matches. If the matches were sufficient for analysis, then a generalized linear model estimated the differences in the probability of funding between the

matched pairs. Using direct matching algorithms, this research examined the statistical necessities to evaluate the ESI policy, finding that sample restriction was necessary to generate a matched sample.

2. Data

The Information for Management, Planning, Analysis and Coordination database (IMPAC II) for NIH applications contains information about funded and unfunded applications that are maintained across time, providing a rich source of longitudinal data about investigators and projects. Because the definition of an ESI is specific to both career stage and prior funding status, applications submitted by ESIs are not directly comparable to applications submitted by experienced investigators (https://grants.nih.gov/policy/new_investigators/index.htm). To address this issue, creating a control group from a cohort of investigators who meet the specifications to qualify for ESI status *prior* to the implementation of the policy was required. To ensure a robust comparison, both demographic and application characteristics were considered to produce comparable propensity for an application to receive funding. Once statistically validated, a comparison between treatment and control groups could evaluate the effectiveness of the policy, providing inferential statistics.

The ESI program specifically targets Research Project Grants (R01) and R01-equivalent applications (refer to https://grants.nih.gov/policy/new_investigators/index.htm#cnaesip). We restricted the pool of applications to new, competing R01 applications that received an impact score during the review process which was used to percentile the application (refer to https://grants.nih.gov/grants/peer-review.htm for information on the peer-review process). R01-equivalent grants were excluded from this research because the specific awards classified as such changes over time and could therefore not be used for comparison purposes across the cohorts (https://grants.nih.gov/grants/glossary.htm#R). The impact score takes into account five review criteria: significance, investigator(s), innovation, approach, and environment. Each category receives a score ranging from 1 (exceptional) to 9 (poor). The mean overall score from each reviewer is then multiplied by 10 and summed as the overall impact score, ranging from 10 to 90. Impact scores are then used to percentile applications. Only a subset of all applications receive percentiles, and as the percentile score was integral to this research, only percentiled applications were included.

The data were further restricted to ICs at NIH that publish both their paylines as well as the added benefit afforded to ESI applications. A payline is a conservative cutoff point where applications scoring below the cutoff point are funded and those scoring above the payline are not funded (Rockey, 2011; NIAID, 2017). One approach ICs used to implement the ESI policy was to create separate paylines for applications submitted by ESIs. For the purpose of this research, six ICs with published overall paylines and ESI-specific paylines for the period analyzed were selected– National Cancer Institute (NCI), National Heart, Lung, and Blood Institute (NHLBI), National Institute on Aging (NIA), National Institute of Allergy and Infectious Disease (NIAID), National Institute of Child Health and Human Development (NICHD), and National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK).[1] Looking at the overall funding at NIH, these six ICs constitute approximately 56 percent of the NIH budget allocated to ICs (HHS, 2016).

[1] The extramural community has resources that identify these payline policies: https://www.einstein.yu.edu/administration/grant-support/nih-paylines.aspx ; and

The treatment group included all R01 applications flagged as ESI submitted between 2011 and 2015. The policy was implemented in 2009; however, the American Recovery and Reinvestment Act of 2009 affected the funding of applications in fiscal years 2009 and 2010. To avoid confounding effects, these years were excluded from the analysis. A corresponding five-year cohort was drawn from ESI-equivalent applications submitted between 2004 and 2008 to form the control group. For the purposes of this research, ESI-equivalent was defined as an application submitted by an investigator aged 42 or under who had no prior substantial NIH independent research awards. In this case, age served as a proxy for career status (https://nexus.od.nih.gov/all/2014/04/29/a-look-at-programs-targeting-new-scientists/). The treatment group was matched directly to the control group using demographic, application, and institutional characteristics.

2.1 Demographic Matching Characteristics

Demographic information is voluntarily entered by the investigator when applying for funding. These data are maintained in the IMPAC II system on the person's profile record. Since this information is voluntary, not all profile records contain demographic information, and therefore cannot be included in the analysis (Ginther et al., 2011; Charette et al., 2016).

Previous research has found that receiving early career stage awards such as training grants, fellowships, and mentored career development programs increases the probability of researchers to successfully transition to independent research careers though the awarding of R01 grants (King et al, 2013; Rangel and Moss, 2004; Wolf, 2002; Zemlo, Garrison, Partridge, and Ley, 2000). While women and other traditionally underrepresented race and ethnic groups have equal if not inflated representation in the pool of these early stage career awardees relative to their representation in the labor market, that is not the case for the group that successfully transitions to independence as measured through successful receipt of an R01 award (Heggeness, Evans, Pohlhaus, and Mills, 2016; Lerchenmueller and Sorenson, 2017). It is possible that the ESI program can increase the representation of these underrepresented groups, thus prior early stage funding, race, ethnicity, and gender need to be included in the matching algorithm.

One of the criterion scores used for determining funding is a score for the investigator(s) on the application. This score specifically looks at the education and training of newly emergent investigators, and as such, the highest degree held by the investigator was also used in the matching algorithm (https://grants.nih.gov/grants/peer/critiques/rpg_D.htm). Including the number of prior attempts captures the applicant's persistence and serves as a control mechanism for career development resulting from feedback. Using the feedback provided by peer reviewers can improve the quality of the application and thus increase the probability of funding (Berger, 2004; Trimble, Bell, Wolf, and Alvarex, 2003). In 2006, NIH began accepting applications from multiple principal investigators (MPIs). In the case of MPI applications, the data are recoded to account for all reported races, ethnicities, the highest degree held, and the maximum number of application submissions.

2.2 Application Matching Characteristics

Several characteristics associated with the application are associated with the probability of receiving funding, such as resubmitting an unfunded application after the initial review, the

https://writedit.wordpress.com/about/nih-paylines-resources/, for example. Additionally, each of these ICs has this information on their public-facing websites.

percentile score of the application, involvement of human subjects, and the IC to which the application was submitted. In one study that specifically examined ESI applications, the National Heart, Lung, and Blood Institute (NHLBI) found that among resubmitted applications, over half of these applications benefitted from the special ESI status (Boyington, Antman, Patel, and Lauer, 2016). Given the increased likelihood of funding for resubmitted applications, and the increased benefit seen by ESI status for resubmitted applications, the submission status was also included in the matching algorithm. The percentile score of the application was found to be the most significant predictor of resubmission (Boyington et al, 2016; Eblen et al., 2016). While the scoring scale has varied over time, the percentile score – the ranking of the application's score within IC among all scored applications – has not, thus warranting inclusion in the algorithm as well.

In addition to the criterion scores, applications including human subjects are subject to additional assessments in peer review associated with human subjects protection and the inclusion of women and racial/ethnic minorities (refer to https://grants.nih.gov/grants/peer/critiques/rpg_D.htm for additional information). Because these applications could differ from those without human subjects, this application characteristic was included in the matching algorithm as well. Another application characteristic included in the matching algorithm was the Institute/Center (IC) to which the application was submitted. Each IC has its own funding guidelines and considerations. Additionally, each IC receives applications specific to the mission of that particular IC, and funding decisions are driven, in part, by the ICs' strategic plans (e.g. https://www.niams.nih.gov/funding/Policies_and_Guidelines/funding_decisions.asp).

2.3 Institution Characteristics

At the institution level, the matching algorithm included the institution type and rank. We computed the institutional ranking as the five-year average rank of each institution based on total overall funding from NIH. The average rank was then recoded to a discrete indicator where a value of 1 was assigned to institutions in the top 25; a value of 2 for those ranked 26 to 50; a value of 3 for those ranked 51 to 100; and a value of 4 for those ranked over 100. An additional indicator at the institution level included the type of institution – medical school, other higher education, or research institute. Institutional characteristics supporting this analysis are based on the IMPAC II institutional profile information collected and maintained with the system, and therefore could differ from the Department of Education's Carnegie classifications.

One of the five criterion scores contributing to the overall impact score of the application is the environment score (refer to https://grants.nih.gov/grants/peer/critiques/rpg_D.htm#rpg_01 for additional information). While this is not the most influential of the scoring criterion (Eblen et al, 2016), it is none the less part of the score used to determine which applications are funded, asking reviewers to consider the institutional role in the probability of the proposed research being successful. The institutional rank, type of institution, and overall resources contribute to the quality of the staff at the institution (Jaffe, 2002; Stephan, 1996; Payne and Siow, 2003; Arora and Gambardella, 2005). Additionally, these same characteristics have a spillover effect on the quality of applications submitted by researchers at that institution (Jacob and Lefgren, 2011; Jaffe et al 1993).

3. Methods

To answer the first research question – can a quasi-experimental design be applied to existing data to infer ESI policy effectiveness – significant exploratory data analysis was conducted prior to

application of modeling to determine the effectiveness of the policy. The initial exploratory data analysis included analyzing descriptive statistics addressing the following data quality arguments as theories that could potentially confound an analysis of the ESI policy:

- ESIs are scored "harder" post-policy implementation than pre-policy implementation;
- Applications submitted post-policy implementation would have been funded regardless of the ESI-specific policies;
- ESIs are better prepared for writing R01 applications post-policy than pre-policy.

While most of the exploratory data analysis relied on analyzing the descriptive statistics of the population, a propensity score model examined the goodness of fit for a matched pairs analysis. These data met the assumptions necessary to apply a propensity model – strongly ignorable treatment assignment and the stable unit treatment value assumption (SUTVA). The first implies that there are no unobservable pretreatment differences between the treatment and the control group (Joffe and Rosenbaum, 1999). We examine the balance between all measured covariates to satisfy this assumption. The latter, SUTVA, has its own assumptions. First, that there is not interference between the treatment and control group. This assumption was met by removing the 35 applications submitted by investigators in both the treatment and control groups. Second, there is only a single version of each treatment. This assumption is met given the guidelines of the policy.

For the purposes of exploratory data analysis, the propensity for each application to be considered an ESI application was calculated. Propensity score models are the conditional probability of treatment (T), in this case being identified as an ESI, given the defined set of characteristics (X):

$$Y(0), Y(1) \perp (T|p(X)) \qquad (Eq.\ 1)$$

where Y is the dichotomous indicator for ESI and X includes gender, race, ethnicity, degree, prior attempts to receive funding (number of applications submitted to NIH), first submission versus resubmissions, human subjects, IC, institutional ranking, and institutional type.

The output from *Equation 1* was used to evaluate the potential matching covariates. The region of common support is the range of the probability of applications being flagged as ESI (Becker and Ichnino, 1999). An ideal control group would have the same distribution of propensity scores as the treatment group. Restricting the data from both groups to the region of common support ensures that any combination of the characteristics used to match the treated case to the control case can occur in both the treatment group and the control group (Bryson, Dorsett, and Purdon, 2002). For the purposes of this research, the minima and maxima comparison technique was used, whereby the data were restricted to the overlapping region of the propensity scores by group.

The propensity score was only used as a form of exploratory data analysis. Treatment cases were matched directly to cases from the control group. The *dist* macro and the *vmatch* macro available in SAS were used to form directly matched pairs.[2] The *dist* macro calculates the weighted Euclidean distance matrix between the treatment case and every control case, based on the specified covariates, using the following equation:

[2] Free SAS code available from Mayo Foundation for Medical Education and Research (MFMER) at http://www.mayo.edu/research/departments-divisions/department-health-sciences-research/division-biomedical-statistics-informatics/software/locally-written-sas-macros

$$D_{ij} = \sqrt{\sum_{i=1}^{d}(x_i - x_j)^2} \qquad (Eq.\ 2)$$

where *i* is the treated case, *j* is the control case, and *x* is the list of specified matching covariates (Gentle, 2007; Larson and Falvo, 2009; Németh & Michaľčonok, 2017). The resulting matrix is an *i* x *j* matrix. The *vmatch* macro evaluates the matrix and selects the best match for each treatment case. In the case of ties, the *vmatch* macro selects *all* cases with the lowest value.

To further evaluate the quality of the matches, the covariate balance and the model sensitivity were tested. Under ideal circumstances, there would not be a statistically significant difference between any of the covariates used to match the treated cases to the control cases. The following equation measured the standardized difference between the prevalence of dichotomous variables in the treatment and control groups:

$$\boldsymbol{d} = \frac{(\boldsymbol{p_{treatment} - p_{control}})}{\sqrt{\frac{\boldsymbol{p_{treatment}}*(\boldsymbol{1 - p_{control}}) + \boldsymbol{p_{control}}*(\boldsymbol{1 - p_{control}})}{\boldsymbol{2}}}} \qquad (Eq.3)$$

where $\boldsymbol{d}$ = difference, and $\boldsymbol{p}$ = prevalence of the dichotomous variable (Austin 2011). Three indicators used in the model were not dichotomous indicators – the number of prior attempts (count), the scored percentile (continuous), and the grouped ranking of the institution (ordinal). The Wilcoxon signed-rank test is a nonparametric test for analyzing matched-pair data with non-normal distributions, which we used to determine the statistically significant difference between these indicators across the two groups (Woolson, 2008).

The final test of the quality of the match was a test for sensitivity using Rosenbaum Bounds based on McNemar's test because the outcome is binary. This test detects the amount of unmeasured bias necessary to change the outcome of the model, making the results no longer statistically valid (Rosenbaum, 2005; Faries et al, 2010). The upper bound calculation is the most salient since the lower bound is always lower than the observed p-value. The upper bound p-value was calculated as follows:

$$\sum_{\boldsymbol{a}}^{\boldsymbol{T}} \binom{\boldsymbol{T}}{\boldsymbol{a}} (\boldsymbol{p}^{+})^{\boldsymbol{a}} (\boldsymbol{1} - \boldsymbol{p}^{+})^{\boldsymbol{T-a}} \qquad (Eq.4)$$

where $\boldsymbol{T}$ is the total number of discordant pairs, $\boldsymbol{a}$ is the number of discordant pairs in which the control case was funded but the treated case was not, and $\boldsymbol{p}^{+}$ is the probability of being exposed accounting for the unobserved confounder (Liu, Kuramoto, and Stuart, 2013).

If the match was deemed acceptable after verifying balanced covariates and the sensitivity of the model, then the difference between the matched pairs can be evaluated to address the second research question – are newly emergent investigators more likely to receive funding for applications submitted post-policy implementation when compared to similar applications submitted pre-policy implementation. To evaluate the difference between the two groups, a generalized linear model estimated the difference in the probability of funding. Generalized linear models include three components – a probability distribution, a linear predictor, and a link function (Dobson and Barnett, 2008).

$$\boldsymbol{E(Y)} = \boldsymbol{\mu} = \boldsymbol{g^{-1}(X\beta)} \qquad (Eq.\ 5)$$

Because the success rate of R01 applications does not exceed 30 percent in any IC, the probability of funding required the use of an overdispersed exponential probability distribution from the binomial family. The derivation and application of the exponential probability distributions, as well as dispersion parameter specifications are not discussed here (see Chapter 3 of Dobson and Barnett, 2008). Instead, the following equation was simplified to show how the probability distribution relates to the other two components of a GLM:

$$\boldsymbol{\mu} = \frac{\mathbf{1}}{\mathbf{1}+\boldsymbol{exp}(-\boldsymbol{X\beta})} \qquad (Eq.\ 6)$$

The linear predictor was tested using the propensity score model outlined in *Equation 1*, therefore $\boldsymbol{X\beta}$ includes the matching covariates mentioned previously. Because we modeled the means directly – in this case probability of being funded for each cohort – the identity link function was applied (Agresti and Finlay, 2009):

$$\boldsymbol{X\beta} = \boldsymbol{ln}\left(\frac{\boldsymbol{\mu}}{\mathbf{1}-\boldsymbol{\mu}}\right) \qquad (Eq.\ 7)$$

The link function transformed the mean to the canonical link or the natural parameter. The model applied to the data was a linear model for a transformed mean of the response variable with a distribution in the binomial exponential family. To ensure the modeled outcome was not affected by the changing economic situation of NIH funding, the estimates were weighted by the annual success rate for unsolicited R01 applications.

As an additional robustness check on the quality of the matches, we ran the models on the full sample, then on three subsamples based on the distribution of the distance indicator for each pair. After the distance (D_{ij}) was calculated using the *dist* and *vmatch* macros, the three subsets included: removing the outliers, the mean and/or median value of D_{ij}, with the final subsample including only those with a distance less than one. Since all other matching covariates are whole numbers, the final subsample consisted of only those pairs that differed by an application percentile score of less than one percent.

Under the assumption that these ESI applications would not have been funded had the ESI-specific funding policy not been in place, the data were recoded such that applications were funded based solely on the published payline of each IC. This approach is crude given that published paylines are estimates produced prior to receiving applications. However, it can serve as a proxy for the benefits afforded to ESI applications under the policy, simulating current funding of ESIs in the absence of the policy.

4. Results

4.1 Descriptive Statistics Exploratory Results

The first two data quality theories focused on the percentiled score of the application relative to the payline. The data used for this analysis does not support either position. Figure 1 shows the percentiled score by cohort, comparing the first year of cohort 1 (e.g. 2004) to the first year of cohort 2 (e.g. 2011). Relative to the control group (FY2004-2008), applications submitted by ESIs between 2013 and 2015 had higher percentile scores than those in the earlier cohort, though none of the differences were statistically significant. Note the range is bound by five percentage points.

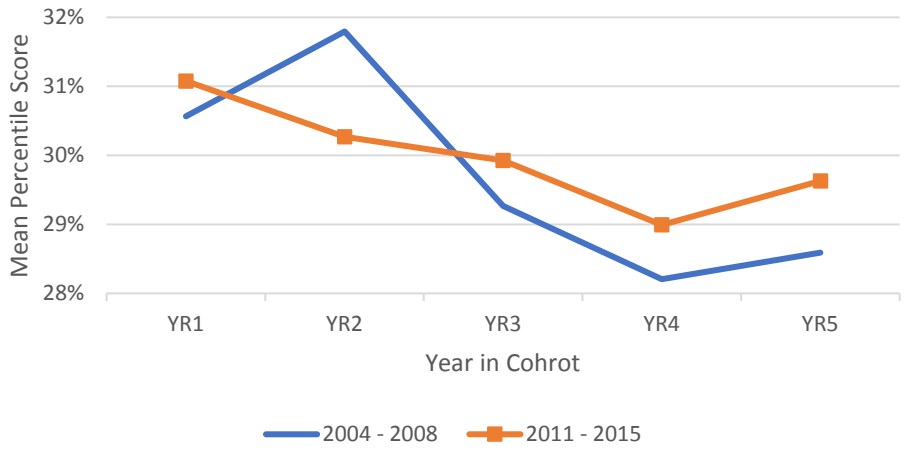

Figure 1: Mean Percentile Score of Applications by Cohort, N=5,954 (2004-2008) and 5,822 (2011-2015).

Figure 2 shows the *funded* ESI applications relative to the IC-specific payline. Of the 1,569 ESI funded applications in sample, over half (50.2 percent to 57.4 percent) benefitted from the ESI-specific funding policy in each fiscal year.

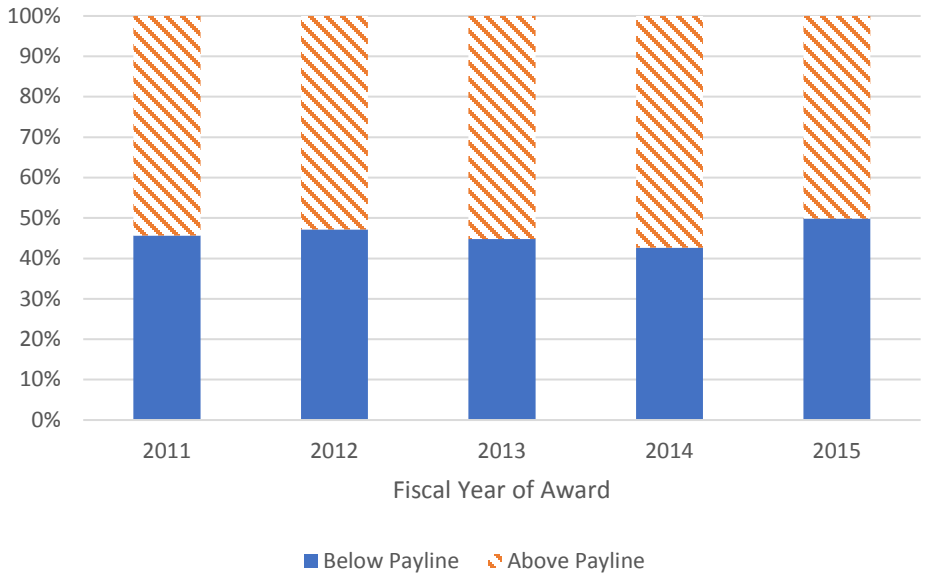

Figure 2: ESI Awards Relative to the IC-Specific Payline, N=723 Below Payline and 846 Above Payline

Figure 3 shows the percent of ESI and ESI-equivalent applicants who received a fellowship (F award), training grant (Trainees), or career development award (K award) prior to applying for the R01 used in this research. While fellowships and training grants have declined slightly over time, there was nearly a 20 percent increase in career development award receipt by ESI and ESI-

equivalent applicants between 2004 and 2015. This trend was not modeled or tested for statistical significance, as it was beyond the scope of this research. Future research should investigate this finding using a time series approach.

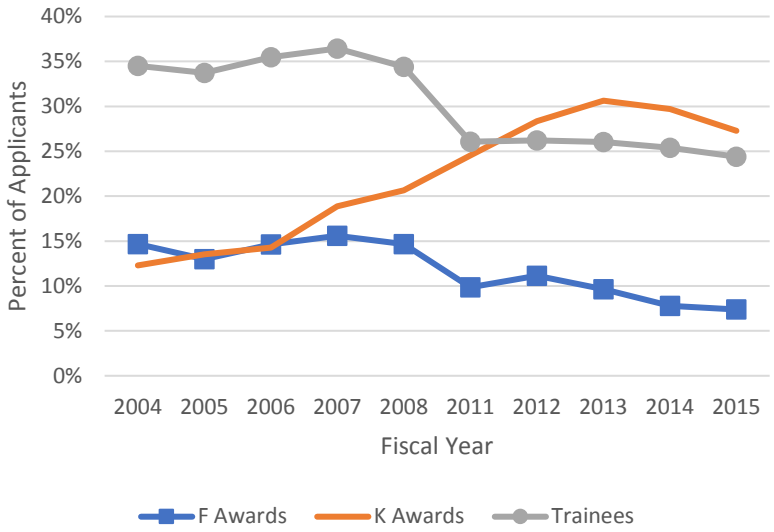

Figure 3: Percent of Applicants with at least One Prior Fellowships, Training Grants, and Career Development Awards, N=11,776.

4.2 Propensity Score Exploratory Results

When restricting to scored applications submitted to ICs with published ESI-specific payline policies, the goodness-of-fit had a p-value of 0.48, which indicates the difference between the two groups was not significantly different from zero. In the control group (FY2004 – 2008), 59 percent of the applications were percentiled, and in the treatment group (FY2011 – 2015) 60 percent of the applications were percentiled. Figure 4 shows the distribution of propensity scores by group.

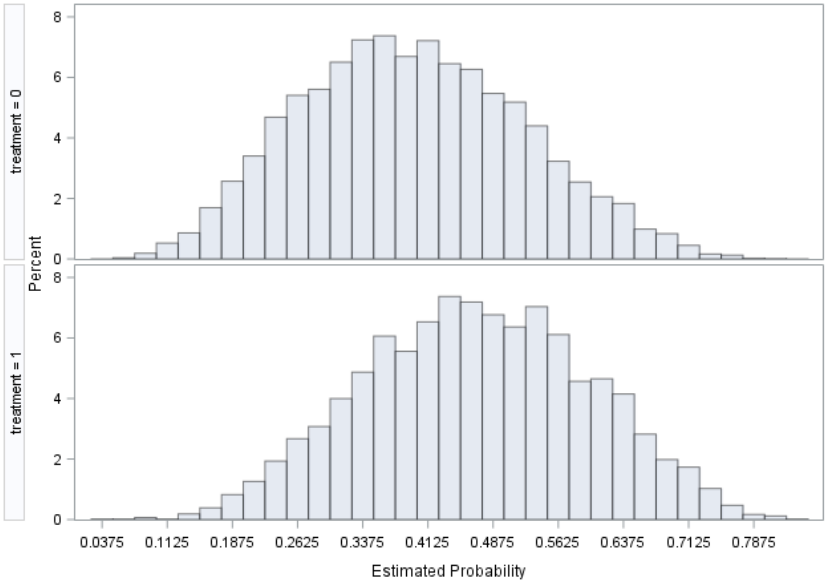

Figure 4: Distribution of Propensity Scores by Group, N=5,954 (2004-2008) and 5,822 (2011-2015).

Figure 5 shows the region of common support for the propensity scores of the two groups. While the two distributions do not align exactly, they are comparable. The region of common support ranged from 0.07 to 0.80, which excluded seven investigators.

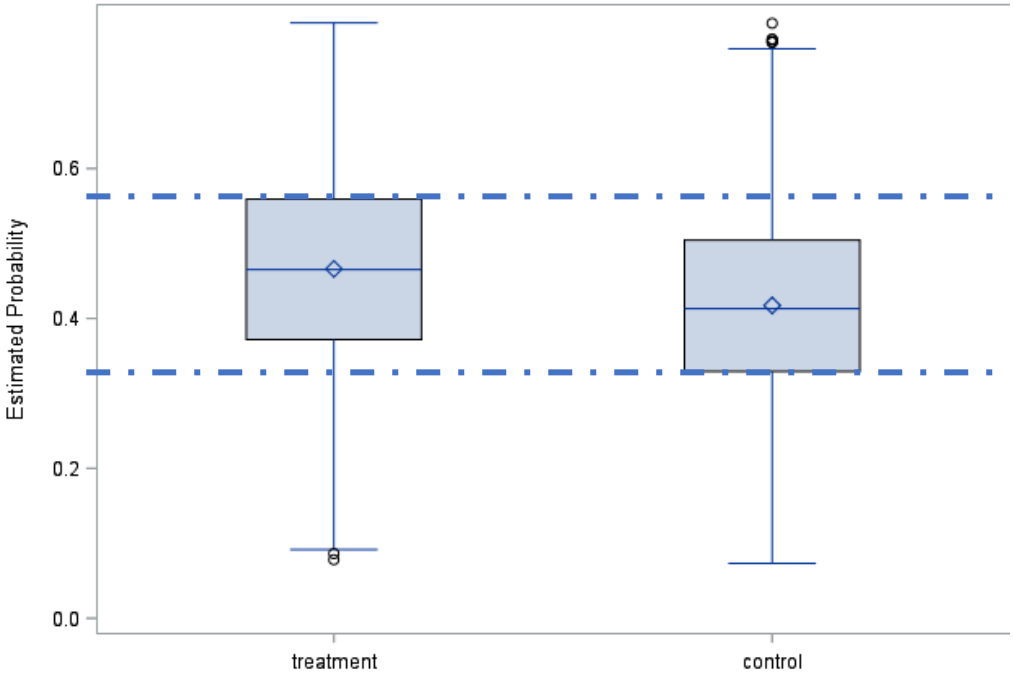

Figure 5: Region of Common Support, N=5,954 (2004-2008) and 5,822 (2011-2015)

All 5,822 applications in the treatment group were successfully matched to at least one of the 8,027 available applications in the control group through the application of the two macros. The macro retained 5,889 treated cases (67 of which had 2 matched control cases) and 5,954 control cases (1 of which was matched to 2 treatment cases), for an overall sample of 11,776 unique applications and a combined 5,955 matched pairs.

The overall distance between all indicators across the treatment and control cases ranged from 0 to 14 for the 5,955 matched pairs (not shown here). Both the mean and median distance (D_{ij})was 3. Of the 20 covariates used to match the treated and control cases, three covariates were not dichotomous indicators – the institutional group ranking, the count for number of applications submitted to NIH, and the application percentile score. Institutions were quartiled, the maximum number of prior applications submitted was 16, and the highest percentiled score was 71. In conjunction with the dichotomous indicators, the potential maximum difference between a treated case and a control case was 107, thus the pair with the greatest distance differed by 13 percent of the relative distance.

Table 1 contains the descriptive statistics by group for the full sample from which the matched cases were drawn. Testing the covariate balance and model sensitivity, we concluded the matched sample was sufficient for analysis, though the results should consider the limitations of the matches. Along with descriptive statistics, Table 1 shows the standardized between all indicators in the matched sample. When examining at the standardized differences, lower values are better and any standardized difference between the two groups that exceeds the 0.10 threshold is a statistically

significant difference. After matching, nine of the included indicators differed significantly between the groups, thus failing to achieve balance.

Table 1: Descriptive Statistics of Matched Pairs

		Control	Treatment	Std. Difference
Person-Level				
	Female	0.31	0.39	0.17*
	Race & Ethnicity (Ref=White)			
	Asian	0.25	0.31	0.15*
	Other Race	0.02	0.04	0.11*
	Hispanic	0.04	0.04	0.04
	Degree (Ref=PhD Only)			
	MD	0.14	0.18	0.11*
	MD-PhD	0.11	0.15	0.11*
	Prior F or T	0.32	0.25	0.16*
	Number of Prior Attempts	1.16	1.12	0.04
Application-Level				
	Resubmission	0.43	0.37	0.12*
	Human Subjects	0.42	0.49	0.14*
	Institute/Center (Ref=NCI)			
	NIAID	0.19	0.15	0.08
	NHLBI	0.21	0.24	0.09
	NIDDK	0.13	0.16	0.11*
	NICHD	0.10	0.10	0.01
	NIA	0.07	0.08	0.04
	Scored Percentile	29.96	30.02	0.06
Institution-Level				
	Institution Type (Ref=Medical School)			
	Higher Education	0.24	0.25	0.01
	Other Institution Type	0.20	0.20	0.00
	Grouped Ranking	2.40	2.36	0.04
N Unique Investigators		5,954	5,822	11,776

***** indicates statistically significant difference between the treatment and control group at the $p \leq 0.1$ level.

The results of the Rosenbaum Bounds sensitivity test are displayed in Table 2. Out of the 5,955 matched pairs, there were 556 discordant pairs, where 342 of these pairs were applications submitted by ESI-equivalents that were funded while the ESI application was not. The McNemar test rejected the null hypothesis that the proportion of funded applications in the treatment group did not differ from the proportion of funded applications in the control group (p-value < 0.001). An unmeasured variable would have to increase the odds of funding by more than 40 percent to negate the treatment effect.

Table 2: Rosenbaum Bounds Sensitivity Test

Gamma	Lower	Upper
1	3.91E-08	0
1.05	1.29E-09	0
1.1	3.72E-11	0.00001
1.15	9.50E-13	0.00011
1.2	2.20E-14	0.00078
1.25	4.44E-16	0.00393
1.3	0	0.01523
1.35	0	0.04678
1.4	0	0.11721*
1.45	0	0.24581
1.5	0	0.44175
1.55	0	0.69535
1.6	0	0.97861

* indicates point where the model is no longer statistically significant

In summary, the descriptive statistics did not show any data quality issues. The propensity modeling suggested matching cohorts across time was possible, and though balance was not achieved, the other propensity modeling assumptions were met. The sensitivity analysis showed strong support for the application of a model to the matched data. Overall, the exploratory data analysis provided support to continue with the policy evaluation.

4.3 Generalized Linear Model Results

Table 3 displays the results of the generalized linear model. The model estimated the control group to be 2.4 percent (p-vale <0.0001) more likely than the treatment group to have an application funded. In other words, all things being equal, the policy has not significantly increased the likelihood of newly emergent investigator applications to be funded.

Table 3: Generalized Linear Model Regressing Probability of Funding on Cohort

	All	$D_{ij} \leq 6$	$D_{ij} \leq 3$	$D_{ij} < 1$
Control	0.299	0.303	0.321	0.286
Treatment	0.275	0.278	0.294	0.260
Difference	0.024***	0.025***	0.027*	0.026
N	11,648	11,397	6,084	522

***p-value ≤ 0.001; ** p-value ≤ 0.01; *p-value ≤ 0.05

Table 4 displays the results of recoding the ESI applications above the IC-specific payline as unfunded and rerunning the generalized linear model. Without the benefit of the ESI-specific payline policy, applications submitted by ESIs were 15.1 percent (p-vale <0.0001) less likely to be funded than applications submitted by ESI-equivalent applicants, controlling for the annual success rates.

Table 4. Generalized Linear Model Regressing Probability of Funding on Cohort[1]

	All	$D_{ij} \leq 6$	$D_{ij} \leq 3$	$D_{ij} < 1$
Control	0.299	0.303	0.321	0.286
Treatment	0.148	0.150	0.160	0.160
Difference	0.151***	0.153***	0.161***	0.126***
N	11,648	11,397	6,084	522

[1]Applications funded as a direct result of the ESI-specific payline policy were recoded as unfunded. ***p-value ≤ 0.001; ** p-value ≤ 0.01; *p-value ≤ 0.05

Both Tables 3 and 4 display the results for the matching validation. The range in differences is minimal – 2.4 to 2.7 percent for the actual model and 12.6 to 16.1 for the simulated model. Additionally, all models had statistically significant differences between the groups with the exception of the most restrictive model regressing the probability of funding on matched pairs that only differed by less than one percentage point in application percentile scores. When recoding for the simulation, 54 percent of newly emergent investigators between fiscal years 2011 and 2015 directly benefitted from the ESI policy and would not have otherwise received funding from NIH.

5. Discussion

With respect to the first research question – is it possible to match NIH grant applications across time – the data quality, propensity model, and the direct match results all indicated sufficient data to match across time. The distribution of the modeled propensity scores were acceptable, and using the minima and maxima criteria for the region of common support only eliminated seven cases. However, despite using the direct matching algorithms and the relative slight differences in distance between the treated and control cases in the matched pairs, there were still statistically significant differences across the groups with respect to the matching covariates, meaning we were unable to achieve covariate balance entirely. The matching did diminish the differences between the groups, and many of the differences were close to no longer being statistically significant after matching. Results from the sensitivity analysis were also promising. It is unlikely that any one indicator could alter the model by 40 percent.

The robustness of the matches were further validated through the application of the model to subsamples. The statistical significance of the findings support the quality of all the matches, even those with the most relative distance between the treated and control cases. Combining the sensitivity analysis with the model validation supports the use of the matching algorithms to compare data groups from different periods of time.

The lack of covariate balance was not the only limitation to this research. A further limitation is the restriction of the sample to the six ICs who publish ESI-payline policies. While this was done intentionally to have a baseline of comparison for future research, this limits the generalizability of the study. Despite these limitations, both the McNemar test and the generalized linear model both showed statistically significant differences between the 2004 – 2008 cohort of ESI-equivalent applications and the 2011 – 2015 ESI applications, such that all things being equal, ESI applications are two percent less likely to receive funding post-policy than pre-policy. This research shows that the deficit is diminished by the policy, though the simulation is not without limitations.

In general, the paylines have been decreasing over time due to the hypercompetitive environment of more investigators competing for fewer funds (Alberts et al, 2014; Kimble et al, 2015). This research shows that the effectiveness of public policies aimed at increasing workforce diversity can be evaluated post-implementation using a pre-policy comparative sample after exploratory data analysis and adjusting for the environment in which the policy is implemented.

6. Conclusions

Matching cohorts across time is difficult and, as was the case in this research, requires severe restrictions to the data included in the sample, thus limiting the generalizability of the study. This study shows that it is imperative to perform robustness and validation checks when using a matching algorithm. Even after restricting the data to scored applications from Institutes/Centers with published ESI-specific payline benefits, while the propensity score distribution appeared sufficient, balance was not achieved between the cohorts. Sensitivity analysis confirmed the model required one specific covariate to alter the probability of being in the treatment group by 40 percent to invalidate the results. The greatest relative distance between a matched treated and control case was 13 percent, and the model results were consistent when restricting to higher quality matches.

The initial model did not provide evidentiary support in favor of the ESI policy. However, when modifying the model to simulate the funding environment in which newly emergent investigators might find themselves if the policy were not in place, the likelihood of funding decreased 15 percent. In other words, modeling policy effectiveness from existing data can be misleading. When comparing the funding pre- and post-policy, it looks as though the policy is not effective. However, after making some assumptions with respect to funding levels, the findings were quite different. The 15 percent deficit was reduced to 2 percent under the current policy. This research shows that 54 percent of newly emergent investigators received funding as a result of the ESI policy. Additionally, this research confirms theories proposed by other researchers (Charette, et al., 2016) that the ESI policy stabilized the proportion of NIH funded ESIs, decreasing the funding deficit experienced by newly emergent investigators.

This research raises additional questions that require future research. The increase in career development awards and the effect of this increase on applying for an R01 grant should be examined. Additionally, while the ESI policy funds newly emergent investigators, this raises the question as to whether the funding is available to sustain these researchers. The NIH issued a new policy, the Next Generation Researchers Initiative (NGRI), which promotes the growth, stability and diversity of the biomedical research workforce. We are continuing to research the later outcomes of investigators funded by the ESI policy, looking specifically at the likelihood of renewing the initial R01 grant as well as the application and funding of subsequent grants.

Acknowledgements

The authors would like to thank Silda Nikaj, Brian, Haugen, Richard Ikeda, and Michael Lauer who provided insight and expertise into this research, as well as comments that greatly improved the manuscript.

References

Alberts, B., Kirschner, M.W., Tilghman, S. and Varmus, H. (2014). Rescuing US biomedical research from its systemic flaws. Proceedings of the National Academy of Sciences of the United States of America (PNAS), 111(16), 5773-5777.

Agresti, A. and Finlay, B. (2009). Statistical Methods for the Social Sciences, Fourth Edition. Pearson Prentice Hall, Upper Saddle River, NJ.

Arora, A. and Gambardella, A. (2005). The impact of NSF support for basic research in economics. Annales d'Economie et de Statistique, 79–80, 91-117

Austin, P.C. (2011). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behavioral Research, 46(3), 399-424.

Basken V, Voosen P (2014) Strapped scientists abandon research and students. Chronicle of Higher Education, 60, 23–23.

Becker, S.O. and Ichino, A. (2002). Estimation of average treatment effects based on propensity scores. The Stata Journal, 2(4), 358-377.

Berg, J. (2010). Scoring Analysis with Funding and Investigator Status. Retrieved June 22, 2017 from https://loop.nigms.nih.gov/2010/09/scoring-analysis-with-funding-and-investigator-status/.

Berger, D.H. (2004). An Introduction to Obtaining Extramural Funding. Journal of Surgical Research, 128(2), 226-231.

Blau, D.M. and Weinberg, B.A. (2017). Why the US science and engineering workforce is aging rapidly. Proceedings of the National Academy of Sciences of the United States of America (PNAS), 114(15), 3879-3884.

Boyington, J.E.A., Antman, M.D., Patel, K.C., and Lauer, M.S. (2016). Toward Independence: Resubmission Rate of Unfunded National Heart, Lung, and Blood Institute R01 Research Grant Applications Among Early Stage investigators. Academic Medicine, 91(4), 556-562.

Bryson, A., R. Dorsett, and S. Purdon (2002). The Use of Propensity Score Matching in the Evaluation of Labour Market Policies. Working Paper No. 4, Department for Work and Pensions.

Charette, M.F., Oh, Y.S., Maric-Bilkan, C., Scott, L.L., Wu, C.C., Eblen, M., Pearson, K., Tolunay, H.E., Galis, Z.S. (2016). Shifting Demographics among Research Project Grant Awardees at the National Heart, Lung, and Blood Institute (NHLBI). Plos One, 11(12): e0168511. Doi: 10.1370/journal.pone.0168511.

Clauset, A., Larremore, D.B., and Sinatra, R. (2017). Data-driven predictions in the science of science. *Science,* 355(6324), 477-480.

Cook I, Grange S, Eyre-Walker A. (2015) Research groups: How big should they be? PeerJ 3:e989 https://doi.org/10.7717/peerj.989

Daniels, R. (2017). A generation at risk: Young investigators and the future of the biomedical workforce. Proceedings of the National Academy of Sciences of the United States of America, 112(2), 313-318.

Dobson, A.J.; Barnett, A.G. (2008). Introduction to Generalized Linear Models (3rd ed.). Boca Raton, FL: Chapman and Hall/CRC.

Dorsey, T. and Wallen, S. (2016) Analysis of NIGMS Funding Rates for Early Stage Investigators and Non-Early Stage New Investigators. Retrieved June 22, 2017 from https://loop.nigms.nih.gov/2016/01/analysis-of-nigms-funding-rates-for-early-stage-investigators-and-non-early-stage-new-investigators/.

Eblen, M.K., Wagner, R.M., RoyChowdhury, D., Patel, K.C., and Pearson, K. (2016). How Criterion Scores Predict the Overall Impact Score and Funding Outcomes for National

Institutes of Health Peer-Reviewed Applications. Plos One, https://doi.org/10.1371/journal.pone.0155060

Faries, D., Leon, A.C., Haro, J.M., Obenchain, R.L. (2010). Analysis of Observational Health Care Data Using SAS. SAS Institute Inc., Cary, North Carolina.

Gentle, J.E. (2007) Matrix Algebra: Theory, Computations, and Applications in Statistics. Springer-Verlag: New York, New York.

Ginther, D.K., Schaffer, W.T., Schnell, J., Masimore, B., Liu, F., Haak, L.L., and Kington, R. (2011). Race, ethnicity, and NIH Research Awards. Science, 333(6045), 1015-1019.

Heggeness, M.L., Evans, L., Pohlhaus, J.R., and Mills, S.L. (2016). Measuring Diversity of the National Institutes of Health-Funded Workforce. Academic Medicine, 91, 1164-1172.

Heggeness, M.L., Carter-Johnson, F., Schaffer, W.T., Rockey, S.J. (2016). Policy Implications of Aging in the NIH-Funded Workforce. Cell Stem Cell, 19(1), 15-18.

Health and Human Services (HHS). (2016). HHS FY2016 Budget in Brief. Retrieved June 22, 2017 from https://www.hhs.gov/about/budget/budget-in-brief/nih/index.html

Jacob, B.A. and Lefgren, L. (2011). The impact of research grant funding on scientific productivity. Journal of Public Economics, 95, 1168–1177.

Jaffe, A.B. (2002). Building programme evaluation into the design of public research-support programmes. Oxford Review of Economic Policy, 18 (1), 22-34

Jaffe A.B., Trajtenberg, R., and Henderson, M. (1993) Geographic localization of knowledge spillovers as evidenced by patent citations. Quarterly Journal of Economics, 108(3), 577-598

Joffe, M.M. and Rosenbaum, P.R. (1999). Invited Commentary: Propensity Scores. American Journal of Epidemiology, 150(4), 327-333.

Jones, B., Reedy, E.J., Weinberg, B.A. (2014). Age and Scientific Genius. National Bureau of Economic Research, Working Paper 19866.

Kimble, J. et al (2015). Strategies from UW-Madison for rescuing biomedical research in the US. eLife 4:e09305.

King, A., Sharma-Crawford, I., Shaaban, A.F., Inge, T.H., Crombleholme, T.M., Warner, B.W., Lovvorn III, H.N., and Keswani, S.G. (2013). The pediatric surgeon's road to research independence: utility of mentor-based National Institutes of Health grants. Journal of Surgical Research, 184, 66-70.

Larson, R. and Falvo, D.C. (2009). Elementary Linear Algebra. Houghton Mifflin Harcourt Publishing Company: Boston, MA.

Lerchenmueller, M.J. and Sorenson, O. (2017). Junior Female Scientists Aren't Getting the Credit They Deserve. Harvard Business Review, Retrieved March 24, 2017 from https://hbr.org/2017/03/research-junior-female-scientists-arent-getting-the-credit-they-deserve.

Levitt, M. and Levitt, J.M. (2017). Future of fundamental discovery in US biomedical research. Proceedings of the National Academy of Sciences of the United States of America, 114(25), 6498-6503.

Matthews, K.R., Calhoun, K.M., Lo, N., and Ho, V. (2011). The Aging of Biomedical Research in the United States. Plos One, 6(12).

Moore, N. (2017). A historical Analysis of NIGMS Early Stage Investigators' Awards Funding. Retrieved June 22, 2017 from https://loop.nigms.nih.gov/2017/04/a-historical-analysis-of-nigms-early-stage-investigators-awards-and-funding/

National Institute of Allergy and Infectious Diseases (NIAID). (2017). Understand Paylines and Percentiles. Retrieved June 22, 2017 from https://www.niaid.nih.gov/grants-contracts/understand-paylines-percentiles.

National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). (2017). Funding Trends and Support of Core Values. Retrieved June 22, 2017 from https://www.niddk.nih.gov/research-funding/funded-grants-grant-history/funding-trends-support-core-values.

National Institutes of Health (NIH). (2017). New and Early Stage Investigator Policies. Retrieved August 22, 2017 from https://grants.nih.gov/policy/new_investigators/index.htm#doni

National Research Council. (2005). Bridges to Independence: Fostering the Independence of New Investigators in Biomedical Research. Washington, DC: The National Academies Press. https://doi.org/10.17226/11249

Németh, M. & Michaľčonok, G. (2017). Preparation and Cluster Analysis of Data from the Industrial Production Process for Failure Prediction. Research Papers Faculty of Materials Science and Technology Slovak University of Technology, 24(39), 111-116. Retrieved 10 Jul. 2017, from doi:10.1515/rput-2016-0024

Payne, A. and Siow, A. (2003). Does federal research funding increase university research output? Advances in Economics and Policy, 3 (1), 1-24.

Rangel, S. and Moss, R.L. (2004). Recent trends in the funding and utilization of NIH career development awards by surgical faculty. Surgery, 136(2), 232-239.

Redelmeier, R.J. and Naylor, C.D. (2016). Changes in Characteristics and Time to Recognition of Medical Scientists Awarded a Nobel Prize. Journal of the American Medical Association, 316(16): 2043-2044. doi:10.1001/jama.2016.15702

Rockey, S. (2011). Paylines, Percentiles and Success Rates. Retrieved June 22, 2017 from https://nexus.od.nih.gov/all/2011/02/15/paylines-percentiles-success-rates/.

Rockey, S. (2012). Age Distribution of NIH Principal Investigators and Medical School Faculty (National Institutes of Health, Bethesda, MD). Retrieved June 21, 2017 from https://nexus.od.nih.gov/all/2012/02/13/age-distribution-of-nih-principal-investigators-and-medical-school-faculty/

Rosenbaum, P.R. (2005). Sensitivity Analysis in Observational Studies. In Encyclopedia of Statistics in Behavioral Science. John Wiley & Sons, LTD.

Stephan, P.E. (1996). The economics of science. Journal of Economic Literature, 34 (3), 1199-1235.

Trimble, E.L., Bell, M., Wolf, J., and Alvarez, R. (2003). Grantsmanship and career development for gynecologic cancer investigators. Cancer, 98(S9), 2075-2081.

Viergever, R.F. and Hendriks, T.C.C. (2015). The 10 largest public and philanthropic funders of health research in the world: what they fund and how they distribute their funds. Health Research Policy and Systems, 14:12 DOI 10.1186/s12961-015-0074-z

Wolf, M. Clinical research career development: the individual perspective. Academic Medicine 2002(77), 1084-1088.

Woolson, R.F. (2008). Wilcoxon signed-rank test. Wiley Encyclopedia of Clinical Trials, Wiley & Sons, Inc. Retrieved April 4, 2017 from http://onlinelibrary.wiley.com/doi/10.1002/9780471462422.eoct979/pdf

Zemlo, T.R., Garrison, T.R., Partridge, N.C., and Ley, T.J. (2000). The physician-scientist: career issues and challenges at the year 2000. FASEB 2000(14), 221-230.