An evaluation of the National Institutes of Health Early Stage Investigator policy: Using existing data to evaluate federal policy

Rachael Walsh*, Robert F. Moore and Jamie Mihoko Doyle

Office of Extramural Research, Statistical Analysis and Reporting Branch, National Institutes of Health, 6705 Rockledge Drive, Rm 4186, Bethesda, MD 20817, USA

*Corresponding author. Email: rachael.walsh@nih.gov

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 who are 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 optimal matching to generate comparable groups pre- and post-policy implementation, generalized linear models were used to evaluate the ESI policy. Due to a lack of control group, existing data from 2004 to 2008 were leveraged to infer causality of the ESI policy effects on the probability of funding applications from 2011 to 2015. This article addresses the statistical necessities of public policy evaluation, finding administrative data can serve as a control group when proper steps are taken to match the samples. Not only did the ESI policy stabilize the proportion of NIH funded newly emergent investigators but also, in the absence of the ESI policy, 54% of newly emergent investigators would not have received funding. This manuscript is important to *Research Evaluation* as a demonstration of ways in which existing data can be modeled to evaluate new policy, in the absence of a control group, forming a quasi-experimental design to infer causality when evaluating federal policy.

Key words: Early Stage Investigator; NIH grant funding; policy evaluation; optimal matching; quasi-experimental design

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 USA (National Research Council 2005). However, current trends in biomedical research funding have created a hypercompetitive environment (Cook, Grange 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 et al. 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 health care 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 et al. 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, long-term consequences to the biomedical research workforce by forcing young scientists to seek 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 the Early Stage Investigator (ESI) policy in 2009, aimed at sustaining a more balanced biomedical research workforce and to lower the increasing age to first major research grant (Heggeness et al. 2016b; Levitt and Levitt 2017). The purpose of the policy was to ‘counter advantages enjoyed by well-established investigators and to encourage early transition to independence’ (NIH 2017). 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 twofold—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 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 purpose of this research was to evaluate the ESI policy, applying new techniques that leverage existing data in the absence of a control group to infer causality. The demonstration of these techniques on an existing policy can be applied to other policy evaluations, providing a new and powerful tool for evaluators to use to enhance the policy process.

2. Theory of evaluation

Nearly three decades ago, Lipsey and Pollard (1989) found less than 10% of evaluations took the fully integrated approach—theory, rationale, and an embedded causal process. Since then, the field of evaluation has grown considerably, and logic models are now incorporated into evaluation plans of federal policy by several countries, including the USA, the UK, and Japan (CDC 1999; HM Treasury 2011; Schröder 2015). The theory and rationale of the ESI policy rely on the assumption that special instructions during the review and funding decision processes would increase the probability of funding applications submitted by newly emergent investigators. However, in accordance with the Federal Grant and Cooperative Agreement Act of 1977, federal policy implementation regarding grant funding is not conducive to an experimental design with a treatment and control group. In the absence of a control group, the policy process is lacking the analytic element of the causal process of evaluation (Grants 2018).

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 causal effects of the policy. For instance, the existing body of literature

relies on descriptive statistics (Dorsey and Wallen 2016; Moore 2017) or was limited to data from one Institute/Center (IC) (Berg 2010; Boyington et al. 2016; National Institute of Diabetes and Digestive and Kidney Diseases (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. Like most federal policy, implementation of the ESI policy was applied to all applications simultaneously, therefore eliminating the potential for a direct comparison control group. However, this research shows evaluators how, under the right constraints, historic data can serve as the control group, providing the analytic element necessary for the causal process.

3. 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. We created a control group from a cohort of investigators who met the specifications to qualify for ESI status *prior* to the implementation of the policy. To ensure a robust comparison, both demographic and application characteristics were considered to produce comparable propensity for an application to receive funding.

For the purposes of this evaluation, we included all new, competing, scored RPG applications flagged for consideration as an ESI application from the following ICs—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) (Einstein College of Medicine 2018; Word Press 2018). These ICs were selected because they publish payline information pertaining to the ESI policy. A payline is a conservative cut-off point where the majority of applications with a percentile rank below the cutoff point are funded and those scoring above the payline are not funded (Rockey 2011; NIAID 2017). One approach of ICs used to implement the ESI policy was to extend the payline for applications submitted by ESIs. Figure 1 illustrates the difference between the overall payline and the ESI payline by fiscal year. The gray dotted line shows the overall payline, while the black dotted line is the ESI extended payline. Thus, the black bars represent the percent of ESIs who would have been funded without the ESI policy benefit (around 15%), while those represented by the shaded portion of the bar are those who were funded as a direct result of the increased paylines from the ESI policy implementation (between 15 and 27%). Each year, between 45 and 50% of funded ESIs would not have been funded without the payline increase from the policy.

The treatment group included 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 5-year cohort was drawn from ESI-equivalent applications submitted between 2004 and 2008 to form the control group. When possible, degree and medical residency

Figure 1. ESI awards relative to the IC-specific payline.

dates were used to flag ESI-equivalent applications in accordance with the ESI policy. When degree or medical residency dates were not provided by the applicant, ESI-equivalent was defined as an application submitted by an investigator of age 42 or under who had no prior substantial NIH-independent research awards. In this case, age served as a proxy for career stage (Rockey 2014), selecting the average age to first major research grant from NIH during the control period (Matthews et al. 2011). The treatment group was matched directly to the control group using demographic, application, and institutional characteristics.

3.1 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. Because more than 12% of the sample were missing either race, gender, or ethnicity, instead dropping these investigators from the analysis as was done in previous research (Ginther et al. 2011; Charette et al. 2016), we retained these investigators and added a dichotomous variable indicating unknown race, ethnicity, and gender.

Previous research has found that receiving early career stage awards such as training grants (T), fellowships (F), and mentored career development programs (K) increases the probability of researchers to successfully transition to independent research careers through the awarding of research project grants (Zemlo et al., 2000; Wolf 2002; Rangel and Moss 2004; King et al 2013). However, when comparing investigators over time, this endogenous characteristic for success has changed between the treatment and control group as evident in Fig. 2. While women and other traditionally underrepresented race and ethnic groups have equal if not inflated representation in the pool of early-stage career awardees relative to their representation in the labor market, it is not the case for the group that successfully transitions to independence as measured through successful receipt of a major research award (Heggeness et al. 2016a; 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.

Figure 2. Percent of applicants with at least one prior fellowship, training grant, and Career Development Award.

One determinant of funding is 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 (NIH 2018). 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 (Trimble et al. 2003; Berger 2004).

3.2 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 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 NHLBI found that among resubmitted applications, over half of these applications benefitted from the special ESI status (Boyington et al. 2016). The percentile score of the application was found to be the most significant predictor of resubmission (Boyington et al. 2016; Eblen et al. 2016). Given the increased likelihood of funding for resubmitted applications, and the increased benefit seen by ESI status for resubmitted applications, the submission status and percentile score were also included in the matching algorithm.

Applications including human subjects are subject to additional assessments in peer review, specifically human subjects protection and inclusion of women and racial/ethnic minorities (refer to https://grants.nih.gov/grants/peer/critiques/rpg_D.htm for additional information). ICs affect peer review as well, and 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). To control for the role of peer review in the application process, applications were flagged as having human subjects and the IC was included.

3.3 Institution characteristics

Grants are funded through the institution with which the investigator is affiliated. At the institution level, the matching algorithm included the institution type and rank. We computed the

institutional ranking as the 5-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 26–50; a value of 3 for those ranked 51–100; 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.

Reviewers are instructed 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 (Stephan 1996; Jaffe 2002; 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 (Jaffe et al. 1993; Jacob and Lefgren 2011).

3.4 Outcomes

The outcome variable of interest for this research is the likelihood of receiving grant support from NIH. For an application to be counted as an award, the application must have a dollar amount greater than 0 and a specified encumbrance date. Because this is a comparison across groups, the aggregate award rate serves as the determinant outcome measure for policy effectiveness.

4. Methods

Significant exploratory data analysis was conducted prior to application of modeling to determine the effectiveness of the policy. We calculated the propensity for each application to be considered an ESI application. Propensity scores are the conditional probability of treatment, in this case being identified as an ESI, given the defined set of characteristics, including gender, race, ethnicity, degree, prior early career stage awards, prior attempts to receive funding (number of applications submitted to NIH), first submission versus resubmissions, percentile score, human subjects, IC, institutional ranking, and institutional type.

The output from the propensity score model was used to evaluate the potential matching covariates. The goodness of fit indicated the difference between the two groups was not significantly different from 0. The region of common support is the range of the probabilities of being flagged as an ESI that overlap between the treatment and control group (Becker and Ichino 2002). 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). We applied the minima and maxima comparison technique, whereby the data were restricted to the overlapping region of the propensity scores by group as shown by the area between the dotted lines in Fig. 3.

Propensity score models often result in an increase in imbalance and model dependence because of the model assumptions (King and Nielsen 2016). The stable unit treatment value assumption (SUTVA) assumes that there is no 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. Additionally, SUTVA assumes there is only a single version

Figure 3. Region of common support.

of each treatment. This assumption is met given the guidelines of the policy. The propensity score approach also assumes that the potential outcomes are orthogonal to, or independent of, the treatment assignment conditioned on the covariates. This assumption, also known as the Strongly Ignorable Treatment Assignment (SITA) assumption, originates from assumptions in randomized control trials (Rosenbaum and Rubin 1983; Joffe and Rosenbaum 1999). We examined the balance between all measured covariates to satisfy this assumption. Unfortunately, the data violated this assumption with the majority of the covariates being significantly different across the treatment and control groups.

These data also violated the parallel trend assumption of quasi-difference-in-difference models (Bertrand, Duflo and Mullainathan 2004). Because we are using a control group from an earlier period, and the applicable population for whom the policy was designed and implemented, the data also violated the assumptions associated with estimating the average treatment effect on the treated, thereby not accurately capturing the differences between the two periods (Imai, King and Stuart 2008). Optimal matching makes no stochastic assumptions. We chose optimal matching over a greedy algorithm, since studies have shown that optimal matching performs as well as, and often better than, greedy algorithms (Rosenbaum 2005; Gu and Rosenbaum 1993).

In optimal matching, cases (i) and controls (j) are matched to minimize the distance (D_{ij}) between the matched pair (Bergstralh and Kosanke 1995; Rosenbaum 2005). The weighted sum of the absolute differences in variables between cases and controls was used to define distances for the optimal match. When higher weights are assigned to a variable, the more likely that a case and control will be matched to that factor (Bergstralh and Kosanke 1995). For this analysis, all variables were weighted equally. The *dist* macro and the *vmatch* macro available in SAS were used to calculate the distance matrix and perform the optimal pair match, respectively (Mayo Foundation 2018). The *dist* macro calculates the difference between the treated case and each of the available control cases. The calculation is based on the specified matching characteristics, to form an $i \times j$ matrix. The *vmatch* macro then processes the matrix, identifying the optimally matched pair. This pair has the smallest D_{ij} value in the matrix. Figure 4 shows the distribution of D_{ij} which ranges from 0 to 6.6, with a mean of 2.3.

To evaluate the quality of the matches, the covariate balance and the model sensitivity were tested. Under ideal circumstances, there

Figure 4. Distribution of distance between matched pairs (D_{ij}).

would not be a statistically significant difference between any of the covariates used to match the treated cases to the control cases, which is why the propensity score was not used. Standardized difference between the prevalence of dichotomous variables in the treatment and control groups was calculated (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). None of the nonparametric tests showed that the difference between the treated indicators and the control indicators was significantly different from 0.

Table 1 contains the descriptive statistics of the matched pairs, including the standardized differences between the group when restricted by D_{ij} . Any standardized difference exceeding 0.10 is indicative of a statistically significant difference between the groups. When retaining the full sample, seven indicators differed significantly. Reducing the difference to two or three matching covariates decreased the number of significantly different indicators to five and one, respectively. When the matched pair is an exact match, or only differs by one of the matching covariates, then there are no statistically significant differences between the groups. However, this reduces the sample size from almost 14,000 to just over 1,000.

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 (Liu, Kuramoto and Stuart 2013). Table 1 also includes the upper bound where the sensitivity model is no longer statistically significant. In other words, when no restrictions are placed on the D_{ij} a given covariate must affect the outcome by more than 40% to nullify the results. However, when D_{ij} is less than or equal to 1, this drops to a mere 5%, meaning the model is highly sensitive.

To evaluate the effects of the ESI policy on newly emergent investigators, a generalized linear model (GLM) estimated the difference in the probability of funding for the treatment and control groups. A weighted GLM with a binomial probability distribution

Table 1. Descriptive statistics

	Proportion		D_{ij}			
	Treated	Control	All	3	2	1
Individual characteristics						
Gender						
Female	0.39	0.30	0.18 ^a	0.14 ^a	0.07	0.04
Unknown	0.01	0.00	0.07	0.03	0.03	NA
Race						
Asian	0.28	0.24	0.09	0.09	0.06	0.03
Other	0.04	0.02	0.11 ^a	0.07	0.02	0.02
Unknown	0.09	0.05	0.15 ^a	0.09	0.05	0.02
Ethnicity						
Hispanic	0.05	0.03	0.08	0.04	0.02	0.06
Unknown	0.12	0.10	0.04	0.01	0.01	0.00
Degree						
MD	0.19	0.15	0.10	0.06	0.04	0.03
MD-PhD	0.15	0.12	0.10	0.08	0.04	0.05
Other	0.00	0.00	0.04	0.04	0.03	NA
Prior Awards						
Prior F	0.08	0.11	0.08	0.07	0.06	0.02
Prior T	0.25	0.31	0.14 ^a	0.11 ^a	0.08	0.01
Prior K	0.26	0.16	0.27 ^a	0.19 ^a	0.11 ^a	0.05
Application characteristics						
Resubmission	0.37	0.45	0.17 ^a	0.13 ^a	0.09	0.05
Human subjects	0.50	0.40	0.19 ^a	0.14 ^a	0.08	0.04
IC						
NCI	0.26	0.31	0.10	0.06	0.05	0.00
NHLBI	0.24	0.21	0.06	0.05	0.04	0.01
NIAID	0.15	0.18	0.09	0.06	0.03	0.01
NIDDK	0.16	0.13	0.07	0.05	0.03	0.00
NICHD	0.11	0.10	0.05	0.03	0.02	0.01
NIA	0.08	0.07	0.05	0.03	0.02	0.00
Institution characteristics						
Higher education	0.26	0.24	0.05	0.04	0.02	0.00
Other	0.20	0.19	0.02	0.01	0.02	0.01
Sensitivity			0.40	0.40	0.25	0.05
N	6,964	6,969	13,933	10,342	5,246	1,252

^aIndicates statistically significant difference between the treatment and control group.

and identity link were used since the outcome is binary, and probability of being funded for each cohort was modeled directly. The dependent variable did not require transformation, as the predicted probability was bounded (0.05–0.83). To ensure the modeled outcome was not affected by the changing economic situation of NIH funding, the estimates were weighted by the applicable fiscal year success rate. The NIH success rate is defined as the percentage of reviewed grant applications that receive funding. Using the fiscal year success rate controls for the increase in overall competitiveness at NIH. We also ran the models using the proportion of researchers fulfilling the requirements for ESI in each fiscal year, controlling for the increase in competition for newly emergent investigators specifically. As an additional robustness check on the quality of the matches, we ran the models on the full sample and then on three subsamples based on the distribution of the distance indicator for each pair as outlined in Table 1.

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 overall payline of each IC. This

approach is crude given that published paylines are estimates produced before 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. After recoding, the GLM was fit to the same subsamples, again weighting by fiscal year success rate and then separately for proportion of applications qualifying as ESI. The difference between the two sets of results then showed us the effects of the ESI policy.

5. Model results

Table 2 displays the results of the GLMs, weighted by fiscal year success rate (the results were similar when weighting by ESI proportion, so they are not reported here). Focusing on the full sample, the model estimated the control group to be 2.4% more likely than the treatment group to have an application funded with the policy in place. The simulation showed that without the policy in place, the gap would be significantly wider at almost 15%. Overall, the ESI policy has increased the probability of funding for ESIs by over 54%, decreasing the deficit in funding by more than 12 percentage points from Cohort 1 to Cohort 2. In other words, all things being equal, the policy has significantly increased the likelihood of newly emergent investigator applications to be funded. Table 2 shows that these results hold when increasing the quality of the match, and the results were all statistically significant, even for the most restrictive model where D_{ij} was less than or equal to 1.

Despite the quality of the matches in the full sample, the conclusions hold. One of the limitations of this research is that this is a simulation. Though the robustness verifications of the modeled results are statistically significant, it is important to keep in mind when interpreting these results that though we are compensating for the lack of a control group through statistical means, this is still a simulation. The sensitivity analyses showed that to nullify the findings, we would need one indicator to affect the probability of funding by 25–40%, meaning it is still possible. From a federal policy evaluation perspective, this is as close to a controlled experiment as we can get given the federal policy parameters and limitations. Without the ESI policy in place, the trend of an increasing age to first major research award would continue.

6. Discussion and conclusions

Overall, this research provides a roadmap for evaluators to fulfill the causal process of evaluation and includes a demonstration of the applicability to a federal policy. Even though federal policy precludes the gold standard experimental design, this research shows that administrative data can be leveraged to infer causality and evaluate policy effectiveness. There are, however, important caveats, and exploratory data analysis is crucial. You must first know your data. This example was specific to NIH-based policy evaluation and focused on the characteristics specific to the NIH grant funding process. However, an initial exploratory analysis will allow this method to be expanded to any policy.

The first step is identifying the matching characteristics that will enable you to have a comparable sample across time. Once you have selected the characteristics relevant to the policy you are evaluating, descriptive analyses will provide you with the ways in which the recoding needs to be applied. The second step is to then evaluate the propensity for treatment of both your treatment and control group,

Table 2. GLM regressing probability of funding on cohort, by match quality and policy implementation simulation

	Control	Treatment	Difference	N
All				
Policy	0.292	0.268	0.024	13, 933
w/o Policy	0.292	0.145	0.147	
Difference		54.1%	–12.3%	
$D_{ij} \leq 3$				
Policy	0.293	0.268	0.025	10, 342
w/o Policy	0.293	0.146	0.147	
Difference		54.5%	–12.2%	
$D_{ij} \leq 2$				
Policy	0.289	0.265	0.024	5, 246
w/o Policy	0.289	0.143	0.146	
Difference		54.0%	–12.2%	
$D_{ij} \leq 1$				
Policy	0.261	0.237	0.024	1, 252
w/o Policy	0.261	0.117	0.145	
Difference		49.4%	–12.1%	

Note: All differences were statistically significant at the $P < 0.001$ level. The simulation is labeled as ‘w/o Policy’.

ensuring your goodness of fit shows no statistically significant differences, and you are restricting the analytic sample to the applicable subset. With respect to the ESI policy evaluation, the difference between the group propensities for treatment was not significantly different from 0 when using the specified covariates, and the analytic sample was restricted to the region of common support, which had considerable overlap across the two groups.

The third step is to evaluate the quality of your matched data set. After selecting the optimal matched pairs, both covariate balance and model sensitivity should be tested. Standardized differences between dichotomous indicators should not exceed the 0.10 threshold, and the difference between nonparametric indicators should not be significantly different from 0. Ideally, none of the matching characteristics would differ between the treatment and the control group. However, additional robustness checks in this example using the ESI policy and grant applications for funding from NIH showed that this does not necessarily have to be the case. We recommend that you apply multiple robustness checks to your data, including a sensitivity analysis, like the Rosenbaum bounds test for sensitivity.

Because the outcome of interest when evaluating the ESI policy was the probability of funding, primal sensitivity analysis was employed instead of simultaneous sensitivity analysis (Liu, Kuramoto and Stuart 2013). Despite the significantly different characteristics between the treatment and control groups, the models were not sensitive to confounding effects when including the majority of the sample. When the sample was restricted such that only one characteristic differed, the proportion of applicants with prior career development awards (K awards), the model sensitivity was still sufficient for analysis. While these mechanisms increase the likelihood of future funding, Heerman, Berg and Barkin (2017) report that only 10–18% of these awardees transition to independence.

The last step is to then fit a model that will evaluate your policy. Exploratory data analysis guided model fit and determined which model specifications needed to be applied to meet the assumptions of the model and the parameters of the data. When comparing data over time, as is usually the case with federal policy evaluations, the

data are dependent upon the environment in which they were collected. Compensatory measures can control for this—for example, weighting the probability of funding by the fiscal year success rates and increasing competition for ESIs specifically. The results from this evaluation show that the ESI policy stabilized the trend of declining success rates for newly emergent investigators, with up to a 12-percentage point increase in the probability of funding. Without the ESI policy, even the most restrictive subsample of matched pairs showed that 49% of newly emergent investigators only received funding as a result of the ESI policy. When the only significant difference between the treatment and control group was the proportion of the sample with a prior career development award, 54% of newly emergent investigators benefitted from the policy implementation and would not have otherwise received funding. While randomized control experiments are still the gold standard, this research shows that it is possible to leverage existing administrative data to infer causality with respect to federal policy.

Acknowledgements

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

Funding

All authors are employees of the National Institutes of Health; in this way, this work was supported by the National Institutes of Health.

References

- Alberts, B. et al. (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: 5773–7.
- 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., and Voosen, P. (2014) 'Strapped Scientists Abandon Research and Students', *Chronicle of Higher Education*, 60: 23.
- Becker, S. O., and Ichino, A. (2002) 'Estimation of Average Treatment Effects Based on Propensity Scores', *The Stata Journal*, 2/4: 358–77.
- Berg, J. (2010) *Scoring Analysis with Funding and Investigator Status*. <<https://loop.nigms.nih.gov/2010/09/scoring-analysis-with-funding-and-investigator-status/>> accessed 22 June 2017.
- Berger, D. H. (2004) 'An Introduction to Obtaining Extramural Funding', *Journal of Surgical Research*, 128/2: 226–31.
- Bergstralh, E. J., and., and Kosanke, J. L. (1995) *Computerized Matching of Cases to Controls*. Technical Report Number 56. <<http://www.mayo.edu/research/documents/biostat-56pdf/doc-10026923>> accessed 25 Sept 2017.
- Bertrand, M., Duflo, E., and Mullainathan, S. (2004) 'How Much Should We Trust Differences-in-Differences Estimates?', *Quarterly Journal of Economics*, 119/1: 249–75.
- 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: 3879–84.
- Boydington, J. E. A. et al. (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: 556–62.
- Bryson, A., Dorsett, R., and Purdon, S. (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. et al. (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.
- Centers for Disease Control and Prevention (CDC) (1999) *Framework for Program Evaluation in Public Health*. MMWR 48 (No. RR-11).
- Clauset, A., Larremore, D. B., and Sinatra, R. (2017) 'Data-Driven Predictions in the Science of Science', *Science*, 355/6324: 477–80.
- Cook, I., Grange, S., and 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: 313–8.
- Dorsey, T., and Wallen, S. (2016) *Analysis of NIGMS Funding Rates for Early Stage Investigators and Non-Early Stage New Investigators*. <<https://loop.nigms.nih.gov/2016/01/analysis-of-nigms-funding-rates-for-early-stage-investigators-and-non-early-stage-new-investigators/>> accessed 22 June 2017.
- Eblen, M. K. et al. (2016) 'How Criterion Scores Predict the Overall Impact Score and Funding Outcomes for National Institutes of Health Peer-Reviewed Applications', *PLoS One*, 11: e0155060. <https://doi.org/10.1371/journal.pone.0155060>.
- Einstein College of Medicine (2018) *NIH Paylines and Success Rates*. <<https://www.einstein.yu.edu/administration/grant-support/ni-h-paylines.aspx>> accessed 31 Jan 2018.
- Faries, D. et al. (2010) *Analysis of Observational Health Care Data Using SAS*. Cary, North Carolina: SAS Institute Inc.
- Ginther, D. K. et al. (2011) 'Race, Ethnicity, and NIH Research Awards', *Science*, 333/6045: 1015–19.
- Grants (2018) *Grant Policies: A Short History of Federal Grant Policy*. <<https://www.grants.gov/web/grants/learn-grants/grant-policies.html>> accessed 2 February 2018.
- Gu, X., and Rosenbaum, P. R. (1993) 'Comparison of Multivariate Matching Methods: Structures, Distances, and Algorithms', *Journal of Computational and Graphical Statistics*, 2: 405–20.
- Heerman, W. J., Berg, J. M., and Barkin, S. L. (2017) 'Mentoring of Early-Stage Investigators When Funding is Tight', *JAMA Pediatrics*, 172: 4–6.
- Heggeness, M. L. et al. (2016a) 'Measuring Diversity of the National Institutes of Health-Funded Workforce', *Academic Medicine*, 91: 1164–72.
- Heggeness, M. L. et al. (2016b) 'Policy Implications of Aging in the NIH-Funded Workforce', *Cell Stem Cell*, 19/1: 15–8.
- HM Treasury (2011) *The Magenta Book: Guidance for Evaluation*. Kew, London: The National Archives. <https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/220542/magenta_book_combined.pdf>.
- Imai, K., King, G., and Stuart, E. A. (2008) 'Misunderstandings between Experimentalists and Observationalists about Causal Inference', *Journal of Royal Statistical Society, Series A (Statistics in Society)*, 171/2: 481–502.
- Jacob, B. A., and Lefgren, L. (2011) 'The Impact of Research Grant Funding on Scientific Productivity', *Journal of Public Economics*, 95: 1168–77.
- 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: 577–98.
- Joffe, M. M., and Rosenbaum, P. R. (1999) 'Invited Commentary: Propensity Scores', *American Journal of Epidemiology*, 150/4: 327–33.
- Jones, B., Reedy, E. J., and Weinberg, B. A. (2014) 'Age and Scientific Genius. National Bureau of Economic Research', Working Paper 19866, The National Bureau of Economic Research, Cambridge, MA.
- King, G., and Nielsen, R. (2016) 'Why Propensity Scores Should Not Be Used for Matching', Working Paper, Harvard University, Cambridge, MA. Copy at <http://j.mp/2ovYGsW>.

- King, A. et al. (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.
- Lerchenmueller, M. J., and Sorenson, O. (2017) 'Junior Female Scientists Aren't Getting the Credit They Deserve', *Harvard Business Review*. <<https://hbr.org/2017/03/research-junior-female-scientists-arent-getting-the-credit-they-deserve>> accessed 24 March 2017.
- 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: 6498–503.
- Lipsey, M., and Pollard, J. (1989) 'Driving toward Theory in Program Evaluation: More Models to Choose From', *Education and Program Planning*, 12/4: 317–28.
- Liu, W., Kuramoto, S. J., and Stuart, E. A. (2013) 'An Introduction to Sensitivity Analysis for Unobserved Confounding in Non-Experimental Prevention Research', *Prevention Science*, 14/6: 570–80.
- Matthews, K. R. et al. (2011) 'The Aging of Biomedical Research in the United States', *PLoS One*, 6/12:e29738.
- Mayo Foundation for Medical Education and Research (2018) *Locally Written SAS Macros*. <<http://www.mayo.edu/research/departments-divisions/department-health-sciences-research/division-biomedical-statistics-informatics/software/locally-written-sas-macros>> accessed 31 Jan 2018.
- Moore, N. (2017) *A historical Analysis of NIGMS Early Stage Investigators' Awards Funding*. <<https://loop.nigms.nih.gov/2017/04/a-historical-analysis-of-nigms-early-stage-investigators-awards-and-funding/>> accessed 22 June 2017.
- National Institute of Allergy and Infectious Diseases (NIAID) (2017) *Understand Paylines and Percentiles*. <<https://www.niaid.nih.gov/grants-contracts/understand-paylines-percentiles>> accessed 22 June 2017.
- National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) (2017) *Funding Trends and Support of Core Values*. <<https://www.niddk.nih.gov/research-funding/funded-grants-grant-history/funding-trends-support-core-values>> accessed 22 June 2017.
- National Institutes of Health (NIH) (2017) New and Early Stage Investigator Policies. <https://grants.nih.gov/policy/new_investigators/index.htm#doni> accessed 22 Aug 2017.
- National Institutes of Health (NIH) (2018) *Definitions of Criteria and Considerations for Research Project Grant Critiques*. <https://grants.nih.gov/grants/peer/critiques/rpg_D.htm> accessed 2 February 2018.
- 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>>.
- 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: 232–9.
- 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: 2043–4.
- Rockey, S. (2011) *Paylines, Percentiles and Success Rates*. <<https://nexus.od.nih.gov/all/2011/02/15/paylines-percentiles-success-rates/>> accessed 22 June 2017.
- Rockey, S. (2012) *Age Distribution of NIH Principal Investigators and Medical School Faculty*. Bethesda, MD: National Institutes of Health. <<https://nexus.od.nih.gov/all/2012/02/13/age-distribution-of-nih-principal-investigators-and-medical-school-faculty/>> accessed 21 June 2017.
- Rockey, S. (2014) *A Look at Programs Targeting New Scientists*. Bethesda, MD: National Institutes of Health. <<https://nexus.od.nih.gov/all/2014/04/29/a-look-at-programs-targeting-new-scientists/>> accessed 21 June 2017.
- Rosenbaum, P. R. (2005) Sensitivity Analysis in Observational Studies. In *Encyclopedia of Statistics in Behavioral Science*. Hoboken, NJ: John Wiley & Sons, LTD.
- Rosenbaum, P., and Rubin, D. (1983) 'The Central Role of the Propensity Score in Observational Studies for Causal Effects', *Biometrika*, 70: 41–55.
- Schröder, D. C. (2015) 'Review of Evaluation: The International Journal of Theory', *Research and Practice*, 11/4: 173–84.
- Stephan, P. E. (1996) 'The Economics of Science', *Journal of Economic Literature*, 34/3: 1199–235.
- Trimble, E. L. et al. (2003) 'Grantsmanship and Career Development for Gynecologic Cancer Investigators', *Cancer*, 98/59: 2075–81.
- 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.
- Wolf, M. (2002) 'Clinical Research Career Development: The Individual Perspective', *Academic Medicine*, 77: 1084–8.
- Woolson, R. F. (2008) Wilcoxon signed-rank test. In *Wiley Encyclopedia of Clinical Trials*. Wiley & Sons, Inc. <<http://onlinelibrary.wiley.com/doi/10.1002/9780471462422.eoct979/pdf>> accessed 4 April 2017.
- Word Press (2018) *Medical Writing, Editing and Grantsmanship*. <<https://writedit.wordpress.com/nih-paylines-resources/>> accessed 31 Jan 2018.
- Zemlo, T. R. et al. (2000) 'The Physician-Scientist: Career Issues and Challenges at the Year 2000', *FASEB*, 14: 221–30.