Understanding the Factors that Modulate the Biomedical Research Workforce

Jingning Li, Xinyi Wang, Xin Yuan , Yunning Zhu
MSBA Candidate, The George Washington University, School of Business
corayuan@gwu.edu

ABSTRACT

As one of the largest centers dedicated to providing biomedical research and training at NIH (National Institutes of Health), NIGMS (National Institute of General Medical Sciences) supports over 3,300 principal investigators per year on R01 research project grants. In order to improve stability of research funding for principle investigators, deliver more efficient budget planning and recognize funding opportunities in the future, NIGMS launched a pilot program to further understand the pool of grantees who was in short funding gaps.

This article identifies five key factors that have the most significant effect on principal investigators' re-entry or exit behavior after gaps and implements a Logistic Regression model for predicting the probability of a principal investigator dropping out of the funding pool in different time frames given a specific date proposed. These findings will help NIGMS estimate the grant application volume, detect the principal investigators who might be at a higher risk of losing funds, and explore efficient ways to allocate biomedical funds for long-term planning purposes. Still, there are constraints need to be overcome for better application and prediction.

KEYWORDS

Biomedical Research Workforce, Research Funding Stability

1. INTRODUCTION

NIH (National Institutes of Health) keeps pursuing fundamental knowledge about the nature and behavior of living systems so as to applying that knowledge to enhance health and reduce illness and disability. To achieve this mission, NIH invests in a research to improve public health. The vast majority of the work in developing and implementing biomedical workforce research programs is conducted by multiple offices, institutes and centers at the NIH.

NIGMS serves as a one of the large institutes supporting for biomedical training at NIH. Annually, it provides more than $2.5 billion in research grants to support fundamental biomedical research projects throughout the US.

For each research project, there is a principle investigator (PI) who will apply for funds from NIGMS to run their projects. With the approvals of the grant application, the project enters the NIGMS funding pool with a specific grant awarded (e.g. R01 or R37). After the grant expires, a renewal application needs to be submitted and approved, or the project could be in a funding lapse ("Gap"). It's very likely for a biomedical research project to experience temporary or permanent funding gaps, since the renewal application may not be approved due to many factors, including the peer review for high scientific and technical merit identification, the breadth and diversity of the Institute's research portfolio, approaches and investigators; the total amount of funding available to the laboratory; and the priority of the research area for the Institute's mission. Sometimes, even applications with outstanding scores might not be funded if the investigator already has substantial other support.

Given previous research from NIGMS, a large portion of principal investigators experience gaps in funding, where the longer the gap is, the less likely for them to return to the NIGMS grantee pool. Gaps in funding could cause an issue for researchers to keep their laboratories running and complete their research, which would also be a great loss for NIGMS. Therefore, preventing funding gaps and improving the stability of biomedical research funding are extremely important for principal investigators, NIGMS, and the scientific research community.

The rest of this paper will further explain the analytics workflow, approaches adapted for variable preparation, model construction and evaluation, and conclusion reached.

2. ANALYTICS WORKFLOW AND TOOLS

The goal of the project is to identify potential factors related to funding gaps and use them to further predict the probability of PI's re-entry or exit the grantee pool.

We used NIH RePORTER as our main data source and pointed out some potential factors, such as stability of funding, inherent competition, objectivity of peer reviews, etc. Considering funding gap as the critical element, seven potential predictors (Gap length, Number of Concurrent Project, Concurrent Funding, Support Year, Fiscal Year, IDeA State, Carnegie Classification of Institutes) were calculated after data cleansing and integration. For dynamic model construction and evaluation, variables were further selected according to their statistical significance and several models were compared based on the prediction accuracy.

The analysis was mainly conducted and programmed with R. Python and SAS/JMP were also used in initial data exploration phase.

3. KEY CONCEPTS AND APPROACHES

3.1 Gap and Gap Length

Four specific dates (01/01/1998, 01/01/2001, 01/01/2006, 01/01/2010) were chosen as the criteria to decide whether a project was in gap on that specific date. The reason for choosing these dates every four years from 01/01/1998 to 01/01/2010 is based on previous research from NIGMS that most gaps last no longer than 4 years.

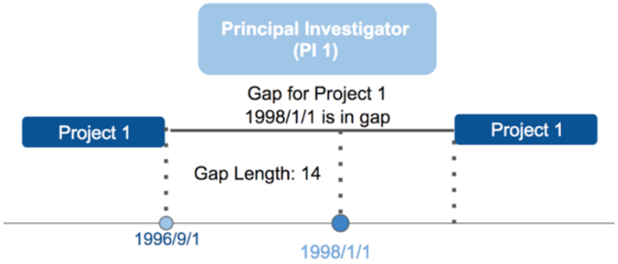
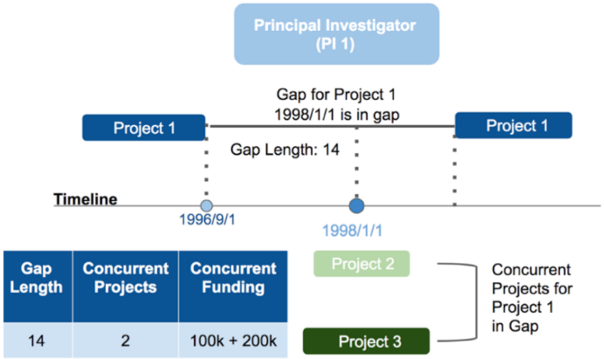


Figure 1.1

Namely, a project (e.g. project 1 in Figure 1.1) will have a funding gap if any time spans between its end dates and the following start dates have these special dates (e.g. 01/01/1998 in Figure 1.1) in the middle. Gap length (e.g. 14 months) was calculated accordingly by getting the time difference between the specific date (e.g. 01/01/1998 in Figure 1.1) detected and the closest end date (e.g. 09/01/1996 in Figure 1.1) before it.

3.2 Number of Concurrent Projects and Concurrent Funding

Two other potential predictors could be generated based on the calculation of gap and gap length in a consistent manner. For each principal investigator, several ongoing projects could run at the same time. It's widely believed that PIs with more concurrent projects running tend to have more motivation of applying for funds. These concurrent projects are usually regarded as financial buffers to help them continue their research with the labs running when one another project is in a gap.



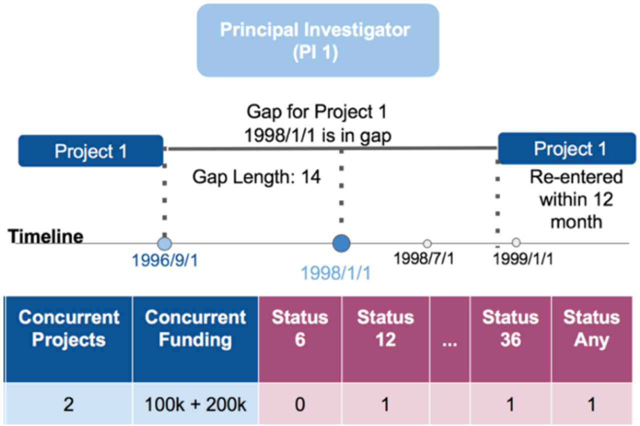
Gap Length	Concurrent Projects	Concurrent Funding
14	2	100k + 200k

Figure 1.2

In the example above, principal investigator 1 (PI 1) has three ongoing projects at the same time, which are project 1, 2 and 3. When project 1 was in a gap on 01/01/1998, project 2 and project 3 were running concurrently on 01/01/1998. In this case, project 2 and project 3 are defined as concurrent projects for project 1 in this specific gap. The value of the concurrent funding equals to the sum of the funding of all the concurrent projects. (e.g. 100k + 200k)

3.3 Dynamic Target Variables

To align with the approach applied to independent variables preparation, a series of target variables for re-entry/exit were measured by the coming six periods of time. Specifically, prediction of re-entry is conducted in a more dynamic way to check if a project would re-enter the funding pool in 6 months, 12 months, 18 months up until 36 months after the specific date on which it was in a gap.



Concurrent Projects	Concurrent Funding	Status 6	Status 12	...	Status 36	Status Any
2	100k + 200k	0	1		1	1

Figure 1.3

In the example above (figure 1.3), project 1 didn't re-enter the funding pool after six months but the re-entry occurred after 12 months. Therefore, the target variable – Status 12, Status 24, Status 36 for project 1 would all be set to 1, whereas Status 6 for project 1 would be 0. For target variable Status Any, if the project ever re-entered the funding pool, it will always be 1, otherwise it would be 0.

3.4 Variables from Complementary Datasets

The Institutional Development Award (IDeA) program broadens the geographic distribution of NIH funding for biomedical research. The program fosters health-related research and enhances the competitiveness of investigators at institutions located in states in which the aggregate success rate for applications to NIH has historically been low. The dummy variable "IDeA State" is a binary variable. If the project's organization state is an IDeA state, the value of this variable will be 1, otherwise it will be 0.

The Carnegie Classification of Institutions of Higher Education is a framework for classifying colleges and universities in the United States. The framework primarily serves educational and research purposes, where it is often important to identify groups of roughly comparable institutions. Each project has a corresponding research organization and a corresponding Carnegie Classification after matching process. ('15' stands for RU/V, which means research university with very high research activity. '16' stands for RU/H, which means research university with high research activity. '25' stands for Spec/Med: Special Focus Institutions – Medical schools and medical centers.) Given the fact that the majority of the research universities in the dataset belongs to '15', we created the variable "Carnegie Classification" as a binary variable, which is 1 or 0 depending on the university is classified as '15' or others ('16', '25').

4. MODEL CONSTRUCTION

The model fitting process started from a basic logistic regression model with Status 12 as the target variable and seven independent variables previously compiled. By test-adding quadratic terms, interactive terms, and checking the degree of statistical significance and strength of relationship, the final best-fitting logistic model was identified gradually.

$$\mathit{Logit}(\mathit{Status12} = 1) =$$

$$68.2788 + 0.0989{*}\mathit{SY} - 0.0031{*}\mathit{SY2} - 0.0829{*}\mathit{Gap.Length} + 0.2451{*}\mathit{Num.Concurrent} - 0.0348{*}\mathit{FY}$$

$$\mathit{Logit}(\mathit{Status24} = 1) =$$

$$47.5375 + 0.1055{*}\mathit{SY} - 0.0033{*}\mathit{SY2} - 0.0745{*}\mathit{Gap.Length} + 0.1515{*}\mathit{Num.Concurrent} - 0.0242{*}\mathit{FY}$$

$$\mathit{Logit}(\mathit{Status36} = 1) =$$

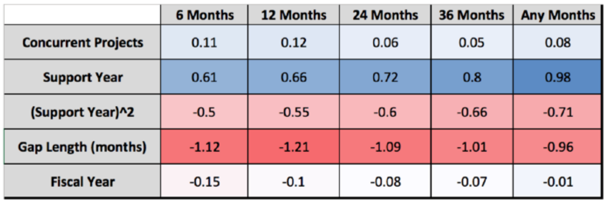
$$40.3159 + 0.1175{*}\mathit{SY} - 0.0036{*}\mathit{SY2} - 0.0707{*}\mathit{Gap.Length} + 0.1316{*}\mathit{Num.Concurrent} - 0.0206{*}\mathit{FY}$$

$$\mathit{Logit}(\mathit{StatusAny} = 1) =$$

$$-3.9713 + 0.1430{*}\mathit{SY} - 0.0039{*}\mathit{SY2} - 0.0690{*}\mathit{Gap.Length} + 0.1795{*}\mathit{Num.Concurrent} - 0.0016{*}\mathit{FY}$$

Figure 1.4

The relative importance of the independent variables after standardization into standard normal distribution: N (0,1), are shown in the Table 1.1:



	6 Months	12 Months	24 Months	36 Months	Any Months
Concurrent Projects	0.11	0.12	0.06	0.05	0.08
Support Year	0.61	0.66	0.72	0.8	0.98
(Support Year)^2	-0.5	-0.55	-0.6	-0.66	-0.71
Gap Length (months)	-1.12	-1.21	-1.09	-1.01	-0.96
Fiscal Year	-0.15	-0.1	-0.08	-0.07	-0.01

Table 1.1

From Table 1.1, Support Year (SY) has the strongest positive effect on re-entry probability, Gap Length has the strongest negative effect on re-entry probability. To specify, controlling for other variables:

- One-unit increase in Gap Length will lead to 8.556% of decrease in the relative probability of re-entry within the next 12 months. A possible explanation for the observed effect is that longer gaps indicate weaker projects that are less likely to be funded again.
- One-unit increase in FY will lead to 3.422% of decrease in the relative probability of re-entry within the next 12 months. It could happen due to the increasing competition for funding.
- One-unit increase in Number of Concurrent Projects will lead to 23.759% of increase in the relative probability of re-entry within the next 12 months. It could be possible when concurrent projects serve as financial buffers.

Another important variable is Support Year (SY), which stands for the age of the project. SY has a quadric effect on re-entry probability, based on the model built, the relationship between SY and relative re-entry probability can be shown as below:

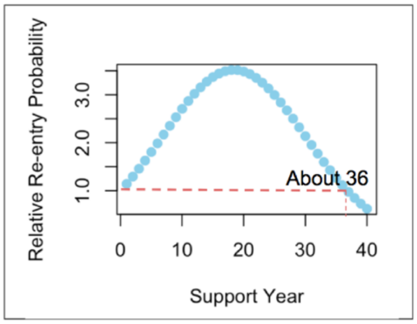


Figure 1.5

Initially, controlling for other variables, relative re-entry probability increases as SY increases. But when SY grows larger than 20 years, the effect of SY on relative re-entry probability will change from positive to negative. A reasonable explanation is that unsuccessful projects tend to end earlier and exit the pool at the beginning, while more successful ones, which have larger number of SYs, are more likely to be funded again and return to the pool. As projects last longer, the researchers become older and may not be able to take charge of their projects anymore. Meanwhile, more new and attractive research topics will be identified, which research teams may transition to.

5. MODEL EVALUATION

After variable preparation and model construction, model validation comes as an integral part of the model development process and would help to show how accurate the chosen model will work in the future. But evaluating model performance with the data used for training is not acceptable in data mining because it can easily generate overly optimistic and over-fit models. In this case, untouched data of 2010 was chosen as a validation dataset (2053 records) to feed to the final model and assessed the performance of model built in the training phase.

The accuracy and performance of final model is validated by comparing model prediction with naïve prediction. Naïve forecasting is an estimating technique in which the last period's actuals are used as this period's forecast, without adjusting them or attempting to establish causal factors.

According to historical records of re-entering during five periods in our training datasets, the re-entering ratios were calculated by dividing the actual number of re-entering projects by the total number of projects and multiplying with the number of records in the validation dataset. As for model prediction, the predicted number of re-entering projects in five periods by the final model with data from validation dataset only. The following plot compares the naïve predictions and model predictions visually:

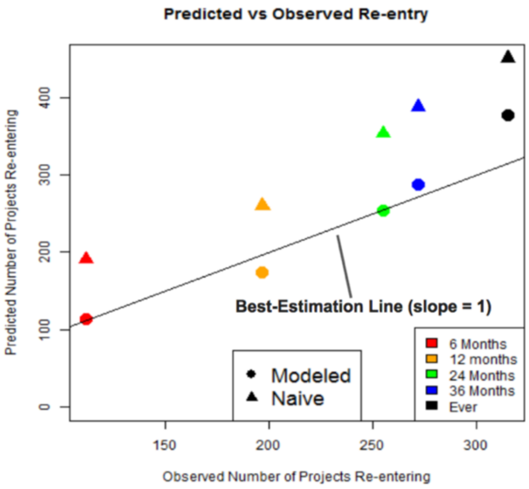


Figure 1.6

In the plot, the X-axis is the observed number of re-entering projects and the Y-axis is the predicted number of re-entering projects. Obviously, there are large distances between triangle dots (naïve predictions) and circle dots (model prediction), which indicates the predictability of this model is significantly different from that of the naïve prediction.

Ideally, the predicted values would equal the observed values, which could be shown as the Best-Estimation Line (slope = 1) in the plot. From Figure 1.5, it's obvious to tell that modeled dots are closer to the Best-Estimation Line and scatter around it, meaning the performance of the final model is much better than that of naive prediction.

6. CONCLUSION

According to data manipulation and model construction, several predictors are evaluated and constituted to the final model. In the order of importance, Gap Length, Support Year, Number of Concurrent Projects and Fiscal Year are key factors, which significantly predict the likelihood that a project will be funded again. Controlling other factors:

- The Larger the Gap Length is, the harder it would be for NIH to sustain projects that have had gaps.
- The relationship between Support Year and re-entry probability depends on the number of Support Years. Up until about support year 20, the relationship is positive; otherwise, it's negative.
- The larger the Fiscal Year is, the lower the probability that a project will re-enter the funding pool.
- The More Concurrent Projects there are, the higher probability there would be for projects that have had gaps to get funded again.

7. LIMITATION & FURTHER APPLICATION

In recent decades, NIH noticed some trends of the US biomedical research workforce, including a decline in the share of key research grants going to younger scientists, a steady rise in the age at which investigators receive their first funding.

Additionally, there are many concerns about the pressures of hyper-competition and the impact of diminishing returns. Some believe that the core problems besetting biomedical research are that too many researchers vying for too few dollars and that people are not paying enough attention to the number of investigators they support. Although scientific research is unpredictable, improving funding stability is still an essential way in support of the future scientific research development.

Given the limited data accessibility and restrained computing capacity of PC, more advanced models and factors could not be discovered. However, the dataset created for this project will be available to other NIGMS researchers who could use the metrics to take the next steps to explore new crucial factors for understanding the biomedical research grantee pool.

Some other directions that would worth considering for further development are:

- Estimating application volume in the future, based on current funded projects and the percentage of the all those expired projects that returned historically.
- Identifying a reasonable time range for inspection. For instance, finding the drop-off rate of those projects that successfully get funded in different periods of time and the length of time for monitoring them.
- Providing insights for generating new predictors with complementary data that only NIH has access to (age, gender, etc.).
- Tagging a "High-Risk Group" with more private data about investigators to better classify them so that NIH can detect which investigators may be at higher risk of losing funding for their projects, as well as designing interventions to keep investigators funded.