Understanding the Factors that Modulate the Biomedical Research Workforce

by Jingning Li, Xinyi Wang, Xin Yuan, Yunning Zhu

M.S. in Business Analytics, Aug 2017, George Washington University

A Thesis submitted to

The Faculty of
the School of Business
of The George Washington University
in partial fulfillment of the requirements
for the degree of Master of Business Analytics

Spring graduation: December 31, 2017

Thesis directed by

Shivraj Kanungo
Chair of Department of Decision Sciences Associate Professor of Decision Sciences
Jacob Basson
Policy Analyst, Statistician at National Institutes of Health
Lisa Hechtman
Program Analyst at the National Institutes of Health
Anna Calcagno
Section Chief at the National Institutes of Health

Abstract of Thesis

Understanding the Factors that Modulate the Biomedical Research Workforce

Getting funded is never easy for members of the biomedical research workforce. NIH has found that when researchers experience time without funding, the longer researchers stay unfunded, the less likely they are to return to the NIH workforce pool.

In the thesis, we dug deeper into NIGMS grants records for key factors to better explain and predict the possibility of investigators losing funding or re-entering the funding pool in different time frames. To do this we conducted variable preparation and logistic regression modeling as well as analysis using R and other analytic tools. The findings will help NIGMS identify better ways to preserve and support researchers, and will help NIGMS estimate the scale of their funding pool in 7 specific upcoming periods so as to efficiently allocate biomedical funds for long-term planning purposes.

Table of Contents

Abstract of Thesis ii

List of Figures iv

List of Tables v

Chapter 1: Introduction 1

Chapter 2: Work Design and Description 4

Chapter 3: Model Construction 11

Chapter 4: Evaluation 18

Chapter 5: Conclusion 22

Reference 25

Appendices 26

List of Figures

Figure 1.1 2

Figure 2.1 4

Figure 2.2 6

Figure 2.3 7

Figure 2.4 8

Figure 3.1 16

Figure 3.2 17

Figure 4.1 19

Figure 4.2 20

Figure 4.3 21

List of Tables

Table 3.1 11

Table 3.2 12

Table 3.3 13

Table 3.4 13

Table 3.5 14

Table 3.6 15

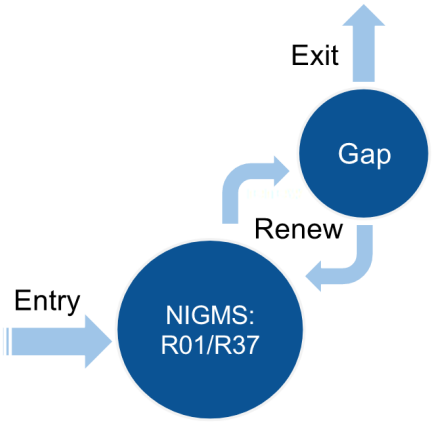
Table 4.1 19

Chapter 1: Introduction

1.1 Background

NIH (National Institutes of Health) keeps pursuing fundamental knowledge about the nature and behavior of living systems and is devoted to applying that knowledge to enhance health, and reduce illness and disability. To achieve this mission, NIH invests in research to improve public health.

The vast majority of the work in developing and implementing biomedical workforce research programs is conducted by 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. And for each project, there is a principle investigator (PI) who will apply funding from NIGMS to run his or her project. Our project will study the behavior of NIGMS PIs and find out factors that influence their behavior.



Graph: NIGMS R01/R37 grantees' flow.

Figure 1.1 NIGMS Funding Workflow

This graph summarizes the project workflow in NIGMS. When an application for funding is approved, the

project enters the funding pool and gets a grant called an R01or R37; after the grant expires, the application needs to be renewed and approved, otherwise the project will exit the funding pool and a gap will start, which means the project experiences time without funding. It's very likely for projects to have gaps since the application may not be approved or the investigator may not re-apply for grants, and a gap could last from a month to several years.

1.2 Problems and Goals

Because the biomedical research workforce must constantly evolve to address new and more challenging health-related questions and problems, an understanding of this particular community's characteristics and dynamics - such as how it navigates the federal funding system - is critical for long-term planning purposes. To achieve that, we will also dig deep into the NIGMS grantee pool which directly reflects the behavior of research workforce: how they enter and exit the pool, how long they stay in the pool and how long the gaps are.

Considering the concept of gaps as well as previous research from NIH, it is not surprising to see that when projects go without funding, the longer the gap is, the less likely for them to return to the NIGMS workforce pool. When a Principal Investigator becomes un-funded, they may not be able to complete their research, or may even have to shut down their lab; also NIGMS will lose its investment when a project cannot be continued, and the science they invested in will not be finished. So preventing funding gaps is therefore extremely important for scientific investigators, NIGMS, and for scientific discovery as a whole. And we consider gap as the most important concept and the main objective we will study for our project.

In order to better maintain the biomedical R&D grant-making ecosystem and develop fit-for-purpose measures/metrics of scientific success and impact, our group will mine data sources, create linkages across datasets, use statistical analyses to help NIGMS efficiently allocate biomedical funds for long-term planning purposes, and identify better ways to preserve and support the research workforce. Specifically, we want to find out how gap length and other potential factors related to gap may affect PIs' behavior and influence the

probability of their projects re-entering the funding pool.

1.3 Resources and Databases

We used NIH RePORTER as our main data source during the program, where almost all the information necessary for the project was available. And we pointed out potential factors such as stability of funding, inherent competition, objectivity of peer review, and transient or permanent exit from NIGMS grantee pool to start our work. For data mining tools, we used mainly R and Excel. Furthermore, we met with our mentors every 2 weeks at their office to discuss and solve the problems together and kept in touch with them about the progress via E-mail.

This is a basic workflow for our project. After downloading original data from RePORTER, we cleaned the raw data by R, extracted useful data that was needed, and stored these data into a new database. Then we identified and developed valuable variables based on the new database and figured out a proper way to form a data table including all the variables that would fit for our further study. That was the most difficult and time-consuming part of our project since it not only required good programming skills but also a thorough and correct interpretation of the whole topic. Finally, we tried multiple statistical analysis models and converted the results from statistical terms into an understandable business case.

1.4 Achievements

As we mentioned before, the term gap means projects temporarily or permanently not funded by NIGMS. And we want to study factors that may affect PIs' likelihood to leave or return to the funding pool after gaps. Until now, we have cleaned data from 1993 to 2015, and identified some potential factors:

- *Gap length*
- *Concurrent project numbers*
- *Concurrent funding*

·*Support year*

·*Fiscal year*

·*Whether the research institution is in an IDeA state*

We chose the best model which gave reasonable explanations after building various models to test the relationship between predictors and target variable, and we analyzed the statistical results combining practical business sense. We found out that some of the predictors did significantly predict PIs' behavior and NIGMS could distribute their funding accordingly.

Our findings are all based on true data and the model has been strongly validated. So NIGMS could trust and actually apply the results in their future work.

Chapter 2: Work Design and Description

2.1 Project Workflow, Methods and Tools

We first detected the most import element: Gaps. Secondly, we generated targets and potential predictors. Thirdly, we used the targets and predictors we generated to construct a model. Several different models have been tried and we chose the most useful and powerful one after comparing them. Then, we analyzed the outcomes of the model and performed the corresponding visualization. Finally, we validated our model, made conclusions and gave some useful recommendations. Figure 2.1 illustrates the workflow interpreted before.

Figure 2.1 Project Workflow

The potential predictors and target variable are very critical, and will be explained later in the following sections. After obtaining the predictors and target variable, we constructed several different models. The primary model is the logistic model, compared with multi-level logistic model. In the process of model construction, we used the stepwise model selection method to select the useful predictors.

We used R to generate potential predictors and target variables. R is a good tool for statistical modeling, so we also used R to construct our model. The model visualization and validation part are also performed in R. Additionally, we used Python and JMP to explore our data and for preliminary model construction.

2.2 Key concepts and Model Design

The whole process of generating potential predictors and target variables, and the model construction is not easy to explain and it is also hard to explain them separately. As a result, we will explain them as a whole process with the corresponding auxiliary pictures by taking a simple example. After explaining the whole process, we will list the potential predictors and target variable and give the exact definition of each to clarify.

We chose four specific dates as the criteria to judge whether one project is in gap on that specific date. They are 01/01/1998, 01/01/2001, 01/01/2006 and 01/01/2010. The reason why we choose these dates every four years from 01/01/1998 to 01/01/2010 is that most gaps last no longer than 4 years.

In this example, we will only look at 01/01/1998. As you can see in this Figure 3.2, project 1 is in gap on 01/01/1998, so project 1 will have one record in our model data.

The first potential predictor we get is gap length, which is calculated by using the specific date we chose: 01/01/1998 minus the project end date of project 1, which in our example is 09/01/1996. So, the gap length for project 1 in our example is 14-month. The reason why we choose the gap length as the potential predictor

is that according to historical funding of NIH, the longer the gap is, the harder the project will come back to the funding pool. The whole process just explained can be illustrated by Figure 2.2.

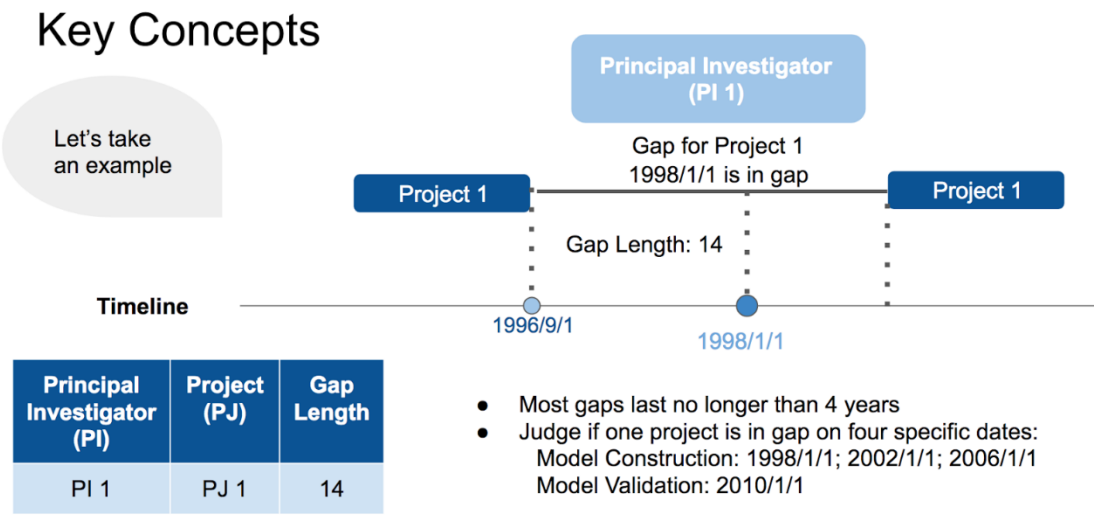
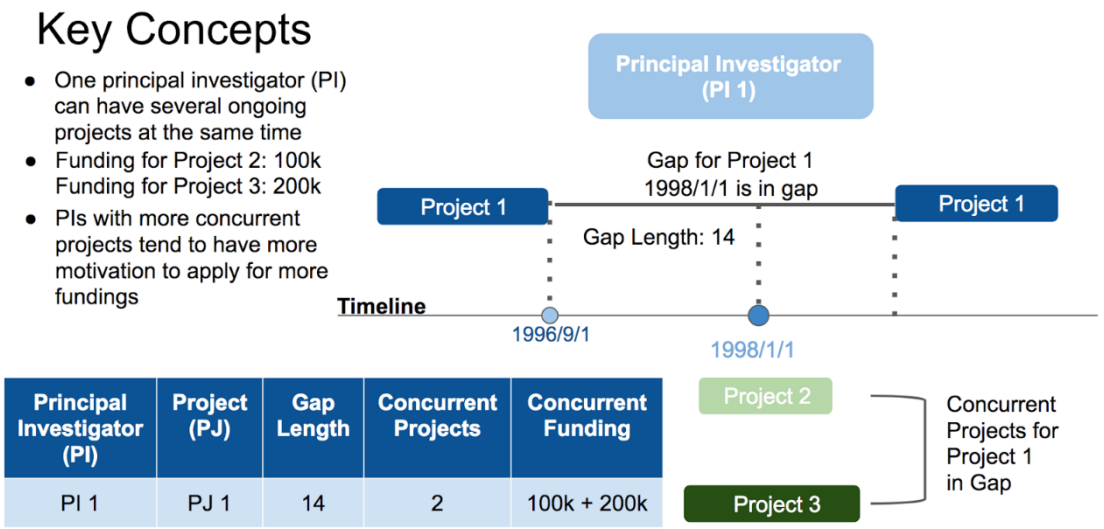


Figure 2.2 Key Concepts I

In addition to gap length, we also considered concurrent number of projects and concurrent funding as two other potential predictors. For most projects, they have one principal investigator we called PI. And for each principal investigator, they may have several on-going projects at the same time. In our example, principal investigator 1 has three ongoing projects at the same time, which are project 1, 2 and 3. When project 1 was in gap on 01/01/1998, project 2 and 3 are ongoing on 01/01/1998. So we considered project 2 and 3 as the concurrent projects for project 1. We also assumed the funding for project 2 and project 3 are 100,000 and 200,000 respectively. So, the concurrent funding for project 1 will be the sum of funding of project 2 and 3, which is 100,000 plus 200,000.

The reason why we choose concurrent projects number and concurrent funding as two potential predictors is that PI with more concurrent projects tend to have more motion to apply for more funding. Besides, PIs with other work going on have a financial buffer. They can keep their research going and their lab running while applying to renew the project in a gap. If a PI only has that one project, they may not be able to continue

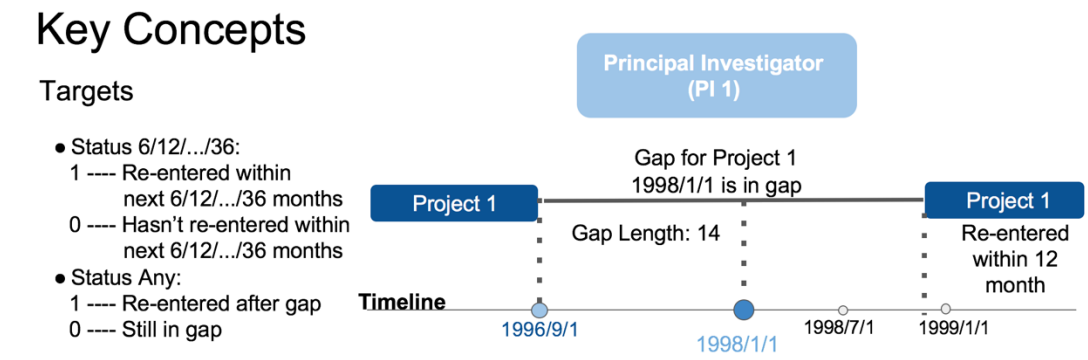
their work (in the very worst case, they could have no money at all - sometimes losing a grant means you have to close your lab). The whole process just explained can be illustrated by Figure 2.3.



Principal Investigator (PI)	Project (PJ)	Gap Length	Concurrent Projects	Concurrent Funding
PI 1	PJ 1	14	2	100k + 200k

Figure 2.3 Key Concepts II

Right now, we will introduce you our target variable. We have several target variables because we want to predict whether a project will re-enter in 6 months, 12 months and 18 months up until 36 months after the specific date on which they are in a gap. In our example, project 1 didn't re-enter after six months but it did re-enter 12 months after that specific date which is 01/01/1998 in our example. So project 1 got the value of 1 for the target variable Status 12 up until to Status 36 but the value of 0 for the target variable Status 6. For target variable Status Any, if the project ever re-entered, it will get the value of 0. The whole process just explained can be illustrated by Figure 2.4.



Principal Investigator (PI)	Project (PJ)	Gap Length	Concurrent Projects	Concurrent Funding	Status 6	Status 12	...	Status 36	Status Any
PI 1	PJ 1	14	2	100k + 200k	0	1		1	1

Figure 2.4 Key Concepts III

2.3 Potential Predictors

As illustrated in the last section, potential predictors are generated by manipulating the metadata. Additionally, a couple of potential predictors are generated by combining an outside source with the metadata. In this section, the exact meaning of each potential predictor, how to generate them and why they are potentially meaningful will be illustrated.

2.3.1 Gap Length

Gap length is the length of the period in which each project is not funded by NIGMS. Gap Length is calculated monthly by subtracting the current project end date and comparing with four specific dates, which are 01/01/1998, 01/01/2001, 01/01/2006 and 01/01/2010, for each project that is in a gap on those four specific dates. As mentioned before, the reason why we think gap length is potentially predictive is that according to historic funding by NIH, the longer the gap, the harder it is for the project to re-enter.

2.3.2 Concurrent Projects

For most projects, there is one corresponding principal investigator (PI). Also, each PI may have several different ongoing projects at the same time. When a project is in a gap at four specific dates, any project which is ongoing at that day will be treated as concurrent project. As a result, the variable Concurrent Projects the number of concurrent projects.

2.2.3 Concurrent Funding

Concurrent Funding is very easy to understand and easy to get after getting the variable Concurrent Projects, which is the total funding of concurrent projects and is obtained by adding all the concurrent projects' funding.

2.3.4 Support Year

Support Year is one column in our metadata. Support Year is the number of years one project has lasted, which is more like the age of project. The reason why we think it may be a useful predictor is that projects with a large support year may be inclined to not re-enter into the funding pool due to large age of PI and a PI's potential interest in new research projects. Additionally, we think Support Year may have an opposite effect at a low number because unsuccessful projects tend to end earlier and exit the pool, while more successful ones, which have larger number of SYs, are more likely to be funded again and return to the pool, we considered adding the quadratic term of Support Year in the model, which will be explained later in the modeling section.

2.3.5 Fiscal Year

Fiscal Year is the government funding year, which is one column in our metadata. It is considered a potential predictor because NIGMS funding policy and circumstances may vary across different fiscal years.

2.3.6 IDeA

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 variable we generated called 'IDeA' is a binary variable. If the project's organization state is an IDeA state, the variable will be appended the value of 1, otherwise it will be appended the value of 0.

2.3.7 Carnegie Classification of Institutions of Higher Education

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, for which we can find a corresponding Carnegie Classification. '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. Due to majority of the research university in our metadata being '15', we modified this variable as the binary variable, 1 stands for the university is classified as 15, 0 means the other way.

2.4 Target Variables

We generated several target variables, which are Status Any, Status 6, Status 12, Status 18, Status 24, Status 30 and Status 36.

2.4.1 Status Any

Status Any is a binary variable which gets the value of 1 if one particular project re-entered after gap and gets the value of 0 if it didn't re-enter after gap.

2.4.2 Status 6/12/18/.../36

Status 6/12/18/.../36 is a binary variable which gets the value of 1 if one particular project re-entered within 6/12/18/.../36 months after gap and gets the value of 0 if it didn't re-enter within 6/12/18/.../36 months after gap.

Chapter 3: Model Construction

3.1 Data Overview

The dataset we used to construct our final models is shown below:

	PPID	pjnum	Gap.Length	SY	FY	idea	Carn15	Status6	Status12	Status18	Status24	Status30	Status36	StatusAny	Num.Concurrent	Funding.Concurrent
1	1857698	39586	40.07671233	6	1993	0	NA	0	0	0	0	0	0	0	0	0.000
2	1857700	36344	9.07397260	9	1994	0	1	0	0	0	0	0	0	0	0	0.000
3	1857700	47922	0.03287671	1	1994	0	1	0	0	0	0	0	0	0	0	0.000
4	1857701	33324	6.08219178	4	1993	0	0	0	0	0	0	0	0	0	0	0.000
5	1858038	48807	40.07671233	1	1994	0	1	1	1	1	1	1	1	1	1	247.149
6	1858040	48449	8.08767123	1	1994	0	1	0	0	0	0	0	0	0	0	0.000
7	1858043	46812	17.06301370	1	1993	1	NA	0	0	0	0	0	0	0	1	154.403
8	1858045	47453	0.03287671	1	1995	0	0	0	0	0	0	0	0	0	0	0.000
9	1858048	42680	5.06301370	5	1993	0	1	0	1	1	1	1	1	1	0	0.000
10	1858048	46749	17.06301370	5	1996	0	1	0	0	0	0	0	0	1	0	0.000

Table 3.1 Data Used for Model Construction

It has 5707 observations and 16 variables. Apart from the potential factors and target variables we mentioned above, we also obtained PPID and pjnum, which represent the ID of each unique applicant and project respectively. In this dataset, each observation stands for one gap record.

3.2 Basic Modeling

To begin, we fit Status12 (whether a project will return within the next 12 months) using all our 7 independent variables: Num.Concurrent, Funding.Concurrent, SY, Gap.Length, idea, FY and Carn15, and logistic modeling, the result of fit is shown in Table 3.2:

Basic Modeling		
Variables	**Estimated Coefficient**	**Significance**
FY	-0.0764	***
Gap.Length	-0.0850	***
SY	0.0271	**
idea	0.1974	
Carn15	-0.1361	
Num.Concurrent	0.4194	
Funding.Concurrent	-3.86E-04	

*Signif.codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '*

Table 3.2 Basic Modeling Result

And the expression of this basic modeling is:

$$Logit(Status12 = 1) \sim FY + Gap.Length + SY + idea + Carn15 + Num.concurrent + Funding.concurrent$$

We could see from above that: FY and Gap.Length significantly negatively predict the odds of a project's re-entry probability and SY has significantly positively predict the odds of a project's re-entry probability.

Deeper investigation reveals that SY (Support Year) has a quadradic, rather than linear relationship with a project's re-entry probability. Based on this knowledge, we added the quadradic form of SY, named it as SY2 and built our second logistic model, with the result of shown in Table 3.3:

Basic Modeling with quadric form		
Variables	**Estimated Coefficient**	**Significance**
FY	-0.0715	**
Gap.Length	-0.0849	***
SY	0.0899	***
SY2	-0.0025	**
idea	0.1740	
Carn15	-0.1316	
Num.Concurrent	0.4126	
Funding.Concurrent	-3.86E-04	

*Signif.codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '*

Table 3.3 Basic Modeling with Quadric Form Result

The expression of this new model is:

$$Logit(Status12 = 1) \sim FY + Gap.Length + SY + SY2 + idea + Carn15 + Num.concurrent + Funding.concurrent$$

Since our new variable, SY2, has a significant negative relationship with re-entry probability, we are confirmed that Support Year does have a quadradic relationship with re-entry probability. According to the new model, FY, Gap.Length, SY and SY2 have significant effect on re-entry probability. However, we also found a strong positive correlation (0.92) between variables Num.Concurrent and Funding.Concurrent, and according to the assumptions of regression, the estimated coefficients and significance may be incorrect. Thus, for further improvement, we fitt Status12 using 5 variables: FY, Gap.Length, SY, SY2 and Num.Concurrent. The result is shown in Table 3.4:

Improved Modeling		
Variables	**Estimated Coefficient**	**Significance**
SY	0.0989	***
SY2	-0.0031	***
Gap.Length	-0.0829	***
Num.Concurrent	0.2451	**
FY	-0.0348	*

*Signif.codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '*

Table 3.4 Improved Modeling Result

The expression of improved modeling is:

$$Logit(Status12 = 1) \sim SY + SY2 + Num.concurrent + FY$$

All five variables have significant effect on re-entry probability, and keeping other variables constant, the increase of SY and Num.Concurrent will increase the re-entry probability, and the increase of SY2, Gap.Length and FY have the opposite effect on re-entry probability.

As we mentioned before, each observation in our dataset stands for one unique gap record, since each applicant may have more than project, and the applicant characteristics, for instance, age, sex, institute, etc, may affect their behavior during gaps, thus affect the re-entry probability of their on-gap project. To eliminate the influence of applicants, we built a multilevel logistic regression model, the upper level of which is PPID, that is the ID for each unique applicant, and the result is shown in Table 3.5:

Multilevel Logistic Modeling		
Variables	**Estimated Coefficient**	**Significance**
SY	0.0989	***
SY2	-0.0031	***
Gap.Length	-0.0829	***
Num.Concurrent	0.2451	**
FY	-0.0348	*

*Signif.codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '*

Table 3.5 Multilevel Logistic Modeling Result

The expression of this multilevel logistic model can be written as:

$$Logit(Status12 = 1) \sim (1 \mid PPID) + SY + SY2 + Gap.Length + Num.Concurrent + FY$$

3.3 Final Modeling

We notice that, the estimated coefficients of this multilevel logistic model are the same as those of simple logistic model. Furthermore, a simple logistic regression model, unlike the multi-level one, allows us to make

predictions using fixed parameters, so in this case, it's a better and more efficient method of modeling. In conclusion, our final models are:

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

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

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

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

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

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

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

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

To find out the relative importance of these variables, we standardized them into standard normal distribution: N (0,1), and generated the following table:

	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 3.6 Coefficients of Standardized Predictors

We can conclude from the above table that variable SY has the strongest positive effect on re-entry probability, and variable 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. One possible explanation of this effect is that longer gaps indicate weaker projects, and weaker projects are less likely to be funded again.

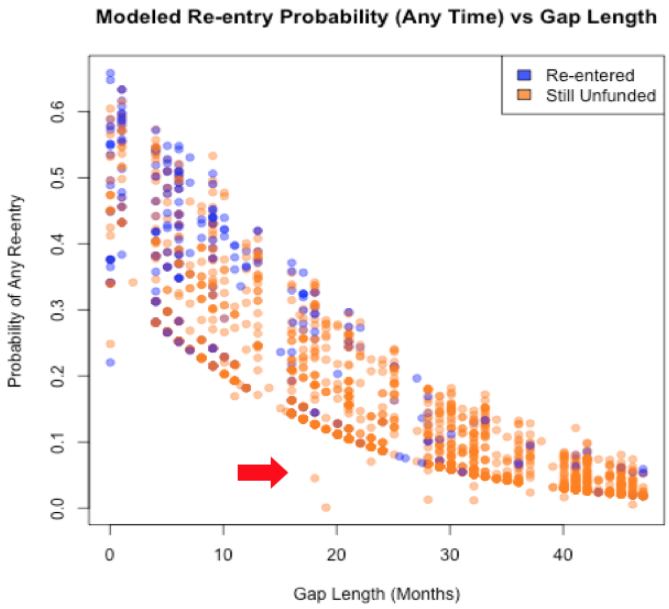


Figure 3.1 The Effect of Gap.Length on Re-entry Probability

To better understand how Gap.Length will affect re-entry probability, we made the above graph, in which blue points stand for projects that really returned after gaps, while yellow points stand for projects that really didn't come back. The X axis of the graph represents the length of gap in term of months, and Y axis represents the re-entry probability of these projects predicted by our model. Take the spot indicated by a red arrow, for example; it has a gap length about 20 months, and our model predicts it has rather low a re-entry probability of about 5%, and in fact, it didn't come back. From this graph we can conclude that, on the one hand, for those projects that did re-enter after their gaps, our model gave a higher prediction of re-entry probability, as we can see that most of the blue points appear on the top of the graph. Thus, at first sight, our model did a pretty good job. On the other hand, the majority of returned projects have a gap length which is no longer than 15 months.

· One-unit increase in FY will lead to 3.422% of decrease in the relative probability of re-entry within the next 12 months. The reason for this is that NIH more strict with applicants in recent years than before, due to increased competition for funding.

· One-unit increase in Num.Concurrent will lead to 23.759% of increase in the relative probability of re-entry within the next 12 months. One possible reason is: investigators with other on-going projects have a financial buffer which could help them keep running their lab while applying to renew the project that's in a gap.

The other important variable is SY, the number of support year, which stands for the age of the project. As we mentioned before, SY has a quadric effect on re-entry probability, based on the model we built, the relationship between SY and relative re-entry probability can be shown as below:

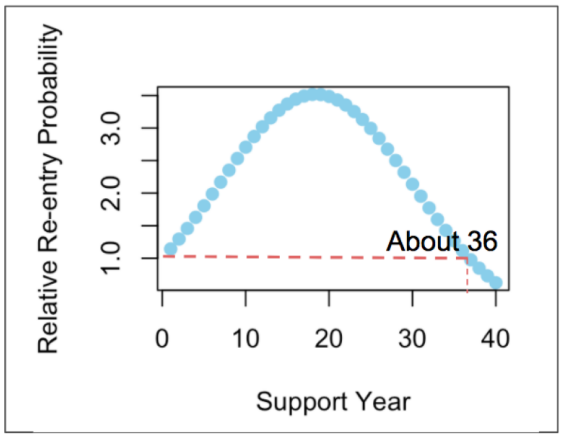


Figure 3.2 Effect of SY on Relative Re-entry Probability

We could see from the above graph that, at first, controlling for other variables, relative re-entry probability increases as SY increases, but as SY grows larger than 20 years, the effect of SY on relative re-entry probability will change from positive to negative. One potential explanation is that at beginning, unsuccessful projects tend to end earlier and exit the pool, while more successful ones, which have larger number of SYs,

are more likely to be funded again and return to the pool. And as the projects last longer, the researchers will grow older and may not be able to take charge of their projects anymore. Furthermore, during the process of these projects, new and more attractive topics will be identified, and the team may transition from the old to the new ones.

A graph of relative re-entry probability by support year can be seen in the appendix.

Chapter 4: Evaluation

4.1 Evaluation of Work

After variable preparation and model construction, model validation comes as an integral part of the model development process and would help to show how well 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. We've already used all data from 1998 to 2010, which are divided every four years as training data sets. In this case, we chose data of 2010 as a validation dataset of 2053 records to feed to the model and assess the performance of model built in the training phase.

4.1.1 Model Validation and Work Comparison

Based on the selection logic we've mentioned in chapter 3, each and every one of records in the validation dataset is selected from the original-downloaded dataset if it's indeed in a gap on 01/01/2010.

Specifically, we kept the actual re-entry status in each of 5 time periods (within 6, 12, 24, 36 months and Any months) as different attributes of each record and applied each of our 5 models (re-entered within 6, 12, 24, 36 months and Any months) to the values in the validation data frame to predict the probability of re-entry in those 5 time frames by using the key predictors we generated initially from the final model. Instead of looking at those probabilities in those 5 time periods separately, we summed them across gaps to get an

estimate of the total number of returning from gaps and compared these to the actual sum of the empirical observation of how many returned historically. Table 4.1 shows the actual re-entering number and predicted re-entering number in different time periods.

	pred	obs
Within 6 Months	113.3831	112
Within 12 Months	174.4660	197
Within 24 Months	253.5946	255
Within 36 Months	287.0928	272
Ever	376.6894	315

Table 4.1 Actual Re-entry vs Predicted Re-entry

As we can see from the table, it's reasonable to say that values in the "pred" column have some kind of relationship with those in the "obs" column. After computing the correlation of "pred" and "obs", we are more confident to say that, since the number 0.97 indicates a strong uphill (positive) linear relationship between the actual re-entery and predicted re-entery in these different time periods. The plot below illustrates the correlation in a better way.

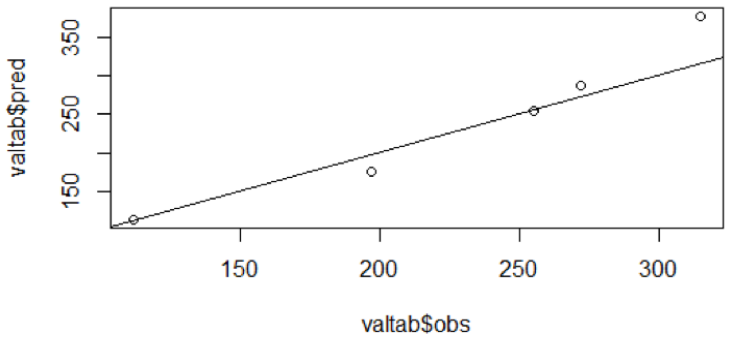


Figure 4.1 The Correlation between Observed and Predicted Re-Entry

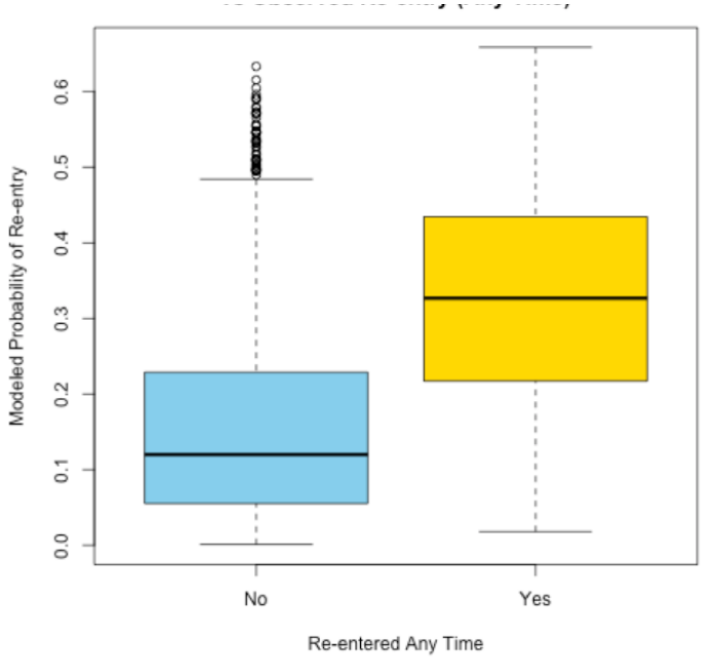


Figure 4.2 Boxplot of Modeled Re-entry Probability (Any Time) VS Observed Re-entry (Any Time)

Additionally, boxplots above also demonstrate that the model behaves as we expected. The model-predicted probability of observed re-entering projects is higher than that of projects which didn't re-enter.

Generally, these few steps above could be regarded as a proof for the bottom-line of the model's predictive accuracy. Furthermore, we validated its predictability and performance 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 5 periods in our training datasets, we calculated the re-entering ratios by dividing the actual number of re-entering projects by the total number of projects and multiplied the re-entering ratios by 2053, which is the number of records in the validation dataset to get the values of naïve prediction. Therefore, it's simply calculated based on the real returning fraction from 1998 to 2006. As for model prediction, which could be told from its name, we predicted the number of re-entering projects in 5

periods by the final model with data from validation dataset only. The following plot shows naïve predictions and model predictions after value-calculation steps demonstrated above.

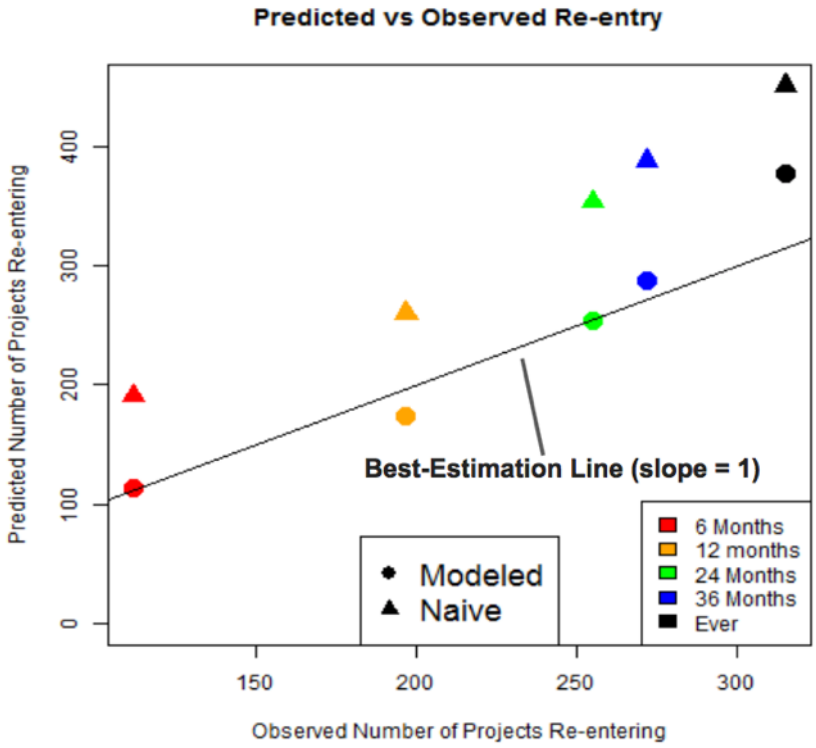


Figure 4.3 Predicted VS Observed Re-entry

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 means the predictability of our model is significantly different from that of the naïve prediction. By common sense, we know the ideal case for prediction and forecasting would be that the predicted values equal the observed values. That's why the Best-Estimation Line shows as the diagonal (slope = 1) in the plot. We can clearly see modeled dots are closer to the Best-Estimation Line and scatter around it. It means the performance of the final model is better than that of naive prediction and proves our final model works well.

4.1.2 Overall Business Evaluation of Work

NIH is one of the largest public funders of biomedical research in the world, investing more than $32 billion a year to enhance life, and reduce illness and disability. Their funded research has led to breakthroughs and new treatments, helping people live longer, healthier lives, and building new areas of knowledge. Once NIH provides fund for a new project, they not only plan to fund it for that year, but have to budget to support that project for many years. Each year NIH has to account for projects they've already committed to funding, but there are also a certain number of projects currently in gaps that will come back in the future. Therefore, knowing how many projects re-enter over time would help NIH set aside funding for the returning projects and efficiently distribute funding across new projects and re-entering projects.

Based on our work that is described in the previous chapters, it's not difficult to estimate the number of returning projects in different time frames by the logistic regression with attributes of current records, and to plan budgets accordingly. Meanwhile, key predictors that are generated from our analysis would provide NIH more characterized information on principal investigators and alerts of potential exiting behavior.

However, with limited data accessibility, only 9 predictors are generated and tested for model construction. It limits the predictive accuracy of the logistic model and leaves more space for development and application extension.

Chapter 5: Conclusion

According to data manipulation and model construction, several predictors are evaluated and constitute 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 for 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

5.1 Unsolved Problems

In recent decades, NIH noticed that a number of distressing trends, including a decline in the share of key research grants going to younger scientists, as well as a steady rise in the age at which investigators receive their first funding, are now a decades-long feature of the US biomedical research workforce. Working committees have proposed recommendations, policy makers have implemented reforms, and yet the trajectory of our funding efforts away from young scientists has only worsened. So are the problems, problems so real that some have gone so far as to write about these concerns. As for these issues, it's less likely to figure out any solutions and breakthrough findings without having access to more private data about principal investigators in the biomedical research workforce, such as an investigator's age.

Additionally, there are many concerns about the pressures of hyper-competition as well as the impact of diminishing returns. Some believe that the core problems besetting biomedical research are "too many researchers vying for too few dollars" and that "we are not paying enough attention to the number of investigators we support, as, given the unpredictable nature of science, we are more likely to generate transformational discoveries by funding more laboratories and research groups." Although scientific research is unpredictable, more useful key factors and models would help NIH better allocate funds. Furthermore, these results are only actionable with more accessibility to more complementary data and time.

5.2 Further Improvement

Taking a step forward to see further developments we could make, we think following directions would be worth considering:

· 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

We are reasonable to believe that these analyses would help NIH better fulfill their financial commitments and sustain the projects they've already invested heavily in. In the long term, it will create a better funding eco-system for the US biomedical research workforce.

References

1. *"A generation at risk: Young investigators and the future of the biomedical workforce"*, Ronald J. Daniels, 313–318, doi: 10.1073/pnas.1418761112

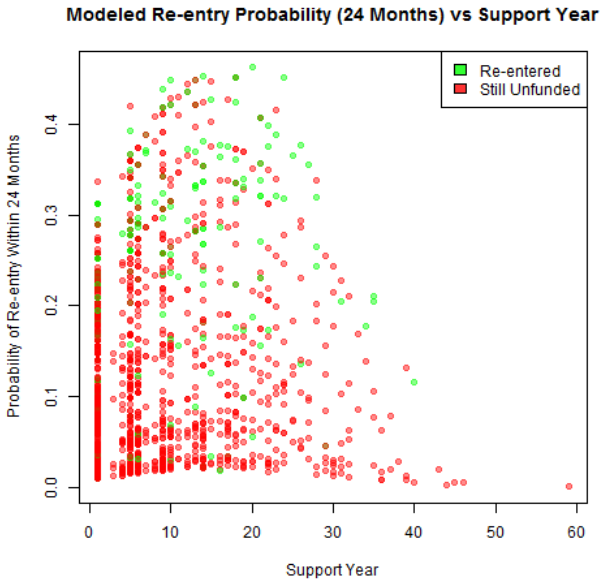
2. *"Implementing Limits on Grant Support to Strengthen the Biomedical Research Workforce*" Mike Lauer, May 2, 2017

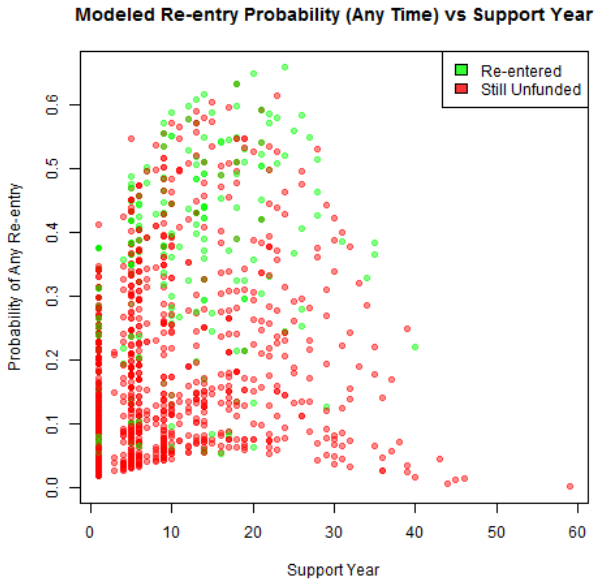
3. *"Institution Development Awards"*, nigms.nih.gov/Research/CRCB/IDeA/Pages/default.aspx, The National Institutes of Health Website

4. *"Types of Application"*, grants.nih.gov/grants/how-to-apply-application-guide/prepare-to-apply-and-register/type-of-applications.htm, The National Institutes of Health, Grants and Funding

Appendices

1. Effect of Support Year on Relative Re-entry Probability





2. Modeled Re-entry Probability vs Observed Re-entry