Replication Appendix for Statistical Analyses in

NIH Funding and the Pursuit of Edge Science

Mikko Packalen and Jay Bhattacharya

This document details the data and programs for replicating the statistical analyses in "NIH Funding funding the Pursuit of Edge Science." The statistical analyses were performed in Stata version 15. Please see SI Appendix for information on the data sources and methods.

INSTALLING FILES

Please first unzip the contents of the file

NIH_Edge_Science_Statistical_Analysis_part1.zip

This creates folder "NIH_Replication" and various subfolders. Then unzip the contents of the files

NIH_Edge_Science_Statistical_Analysis_part2.zip
NIH_Edge_Science_Statistical_Analysis_part3.zip
NIH_Edge_Science_Statistical_Analysis_part4.zip
NIH_Edge_Science_Statistical_Analysis_part5.zip

to folder "NIH_Replication\IdeaInputFiles".

FOLDERS

\NIH_Replication\	.do files .dta files that link datasets .dta file for Table S8
\NIH_Replication\Characteristics	Characteristics .dta files
\NIH_Replication\IdeaInputFiles	Idea input .dta files
\NIH_Replication\Results	All programs print results here
\NIH_Replication\Temp	Intermediate .dta files generated by programs

	for Tables S1-S2 and Table S7 (The .do files generate these files but the files are also included)
\NIH_Replication\Temp2	Intermediate .dta files generated by programs for Figures 2-4 and Figures S1-S3 (The .do files generate these files but they are also included)

VARIABLE NAMES AND DESCRIPTIONS

Variable name Description

pmid	Pubmed ID of a published research article
year	Year of publication of research article
NIH_status	NIH Funding status (1 for papers with NIH Funding, 0 otherwise)
journalcategory	NLM journal category name (125 categories); proxy for research area
journalcategoryid	Numeric representation of NLM journal category
categoryname	UMLS category of UMLS term (127 categories); proxy for idea type
categoryset_number	Numeric representation of UMLS category
categorygroupname	UMLS category group (we linked each UMLS category to one of *Basic Science*, *Clinical*, and *Miscellaneous*)
word	The newest term linked to a given UMLS category
cohort	Cohort of the newest term linked to a given UMLS category

DATASETS

1. Data on idea inputs

File names: \IdeaInputs\'year'\Newest_ideainput_`ideatype_id'.dta

- 'year' has range 1990-2016
- `ideatype_id' has range 0-127; `ideatype_id'=0 captures newest idea input across all idea types
- Variables: pmid, categoryset_number, word, cohort
- Observations at the (pmid, idea type) level
- Includes newest UMLS term and cohort for every (pmid, idea type) combination

2. Data on other research article characteristics

File names: \Characteristics\characteristics_`analysis_type'.dta

- There are 6 files; `analysis_type' has 6 values
- 5 characteristics_`analysis_type'.dta files represent a time period
  - 2010-2016; file: `analysis_type'=93
  - 1990-1999; file: `analysis_type'=90
  - 2000-2009; file: `analysis_type'=91
  - 2010-2014; file: `analysis_type'=88
  - 2015-2016; file: `analysis_type'=89
- The file characteristics_`analysis_type'.dta with `analysis_type'=66, research area is determined based on the MeSH vocabulary (in other files research area is determined based on NLM journal category). In this file, the time period is 2010-2016. For ease of analysis, also for this file the research area variable is labelled as "journalcategoryid".
- Variables: pmid, journalcategoryid, year, NIH_status
- Observations at the (pmid, journal category) level

3. Data linking categoryset_number to categoryname and categorygroupname

File name: linkfile_for_ideatypeid_to_ideatypename.dta

- Variables: categoryset_number, categoryname, categorygroupname
- Links each categoryset_number with categoryname and categorygroupname

4. Data linking journalcategory_id to journalcategory_name

File name: linkfile_for_journalcategoryid_to_journalcategoryname.dta

- Variables: journalcategoryid, journalcategory
- Links each journalcategoryid with journalcategory

5. Lists of selected idea types and selected journal categories

File name: selected_ideatypes_for_TABLE_S1.dta

- List of those idea types that are represented in Table S1
- Variables: categoryset_number

File name: selected_journalcategories_for_TABLE_S2.dta

- List of those journal categories that are represented in Table S2
- Variables: journalcategoryid

REPLICATION OF FIGURE 1

Run `draw_funding_by_cohort_TABLE_1.do`

As inputs this program uses the characteristics file \Characteristics\characteristics_93.dta and idea input files \IdeaInputs\'year'\newest_ideainput_`ideatype_id'.dta from years 2010-2016.

REPLICATION OF FIGURES 2-4 AND FIGURES S1-S3

First run `transform_to_ideatype_journalcategory_level.do`

This program transforms the (pmid, idea type, journal category) level data to (year, idea type, journal category) level data.

As inputs this program uses the characteristics files \Characteristics\characteristics_`analysistype'.dta (with `analysistype'=93, 90, 91, 88, 89, 66) and idea input files \IdeaInputs\'year'\newest_ideainput_`ideatype_id'.dta from years 1990-2016.

Next, run `draw_all_figures_FIGURES_2_to_S3.do`

As inputs, this program uses the files generated by the above program `transform_to_ideatype_journalcategory_level.dta` and the link file linkfile_for_ideatypeid_to_ideatypename.dta that includes the idea type group (*Basic Science*, *Clinical*, or *Miscellaneous*) of each idea type.

REPLICATION OF TABLE S1

Run `calculate_edgefundingratios_by_ideatype_TABLE_S1.do`

As inputs this program uses the characteristics files \Characteristics\characteristics_`analysistype'.dta (with analysistype=93, 90, 91, 88, 89) and idea input files \IdeaInputs\'year'\newest_ideainput_`ideatype_id'.dta from years 1990-2016, as well as the link file linkfile_for_ideatypeid_to_ideatypename.dta and the file selected_ideatypes_for_TABLE_S1.dta that selects the individual ideatypes that are shown in the table.

REPLICATION OF TABLE S2

Run `calculate_edgefundingratios_by_journalcategory_TABLE_S2.do`

As inputs this program uses the characteristics files \Characteristics\characteristics_93'.dta and idea input files \IdeaInputs\'year'\newest_ideainput_`ideatype_id'.dta from years 2010-2016, as well as the link file linkfile_for_journalcategoryid_to_journalcategoryname.dta and the file selected_journalcategories_for_TABLE_S2.dta that selects the individual journal categories that are shown in the table.

REPLICATION OF TABLE S3

Run `construct_ideatype_list_TABLE_S3.do`

As inputs this program uses the characteristics file \Characteristics\characteristics_93.dta and idea input files \IdeaInputs\'year'\newest_ideainput_`ideatype_id'.dta from years 2010-2016, as well as link file linkfile_for_ideatypeid_to_ideatypename.dta

REPLICATION OF TABLE S4

Run `construct_ideainput_list_TABLE_S4.do`

As inputs this program uses the idea input files \IdeaInputs\'year'\newest_ideainput_`ideatype_id'.dta from years 2010-2016, as well as link file linkfile_for_ideatypeid_to_ideatypename.dta

REPLICATION OF TABLE S5

Run `construct_journalcategory_list_TABLE_S5.do`

As inputs this program uses the characteristics file \Characteristics\characteristics_93.dta and idea input files \IdeaInputs\'year'\newest_ideainput_`ideatype_id'.dta from years 2010-2016, as well as link file linkfile_for_journalcategoryid_to_journalcategoryname.dta

REPLICATION OF TABLE S6

Run `construct_cohort_distribution_TABLE_S6.do`

As inputs this program uses the characteristics file \Characteristics\characteristics_93.dta and idea input files \IdeaInputs\'year'\newest_ideainput_`ideatype_id'.dta from years 2010-2016.

REPLICATION OF TABLE S7

Run `test_differential_support_over_time_TABLE_S7.do`

As inputs this program uses characteristics files \Characteristics\characteristics_90.dta, \Characteristics\characteristics_91.dta, and \Characteristics\characteristics_93.dta and idea input files \IdeaInputs\'year'\Newest_ideainput_`ideatype_id'.dta from years 1990-2016.

REPLICATION OF TABLE S8

Run `construct_nonNIH_funding_sources_TABLE_S8.do`

As input, this program uses the file non_NIH_funding_sample_150.dta.