Indirect Cost Recovery Rates: A New Dataset

Pierre Azoulay (MIT and NBER) Andrew Breazeale (CMU)
Bhaven Sampat (Columbia University and NBER)*

September 2019

Abstract

About a quarter of U.S. research funding is paid as indirect costs. Accordingly, indirect cost rates (ICR) and policy have been the subject of considerable debate. A major constraint on previous research has been a lack of good longitudinal data for a large number of institutions; numerous Congressional and other government reports have called for better data. In this paper, we describe the construction of a new panel dataset of negotiated indirect cost recovery rates by institution, based on information obtained through a series of Freedom of Information Act (FOIA) requests. We also provide links to both the raw and processed data.

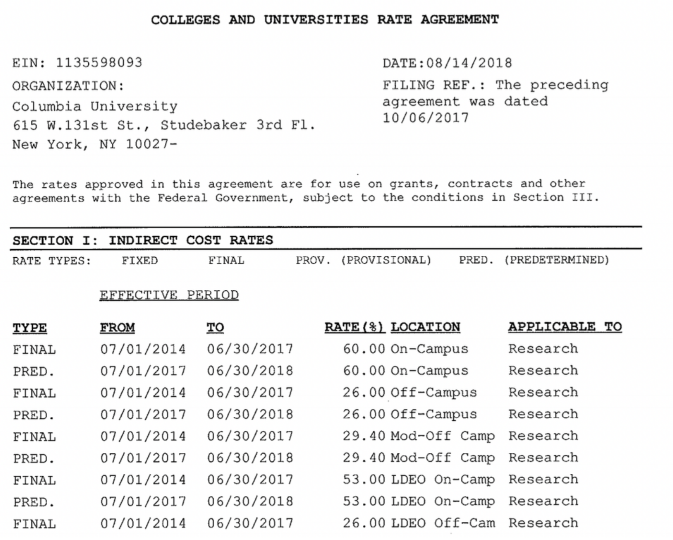
1 Introduction

Institutions (e.g., universities, research hospitals, and independent research institutes) negotiate their indirect cost rates with a cognizant federal agency (i.e., the one that awards the majority of the institution's funding). For academic institutions, this is typically the Department of Health and Human Services. Each institution's current negotiated indirect cost rate agreement (NICRA) is generally made public. For example, Figure 1 shows an excerpt from Columbia University's current rate agreement.[1]

Though current rate agreements often are publicly available, few institutions report historical rates. This makes longitudinal analysis difficult. In this paper we describe the construction of new longitudinal datasets of indirect cost rates by institution. We obtained raw rate information using data from two FOIA requests:

*This research was supported by NSF Award 1735413 "The Effects of Federal Indirect Cost Recovery on Academic Institutions"

[1] This is available from: `https://finance.columbia.edu/content/correct-fa-costs`.



COLLEGES AND UNIVERSITIES RATE AGREEMENT

EIN: 1135598093
ORGANIZATION:
Columbia University
615 W.131st St., Studebaker 3rd Fl.
New York, NY 10027-

DATE:08/14/2018
FILING REF.: The preceding
agreement was dated
10/06/2017

The rates approved in this agreement are for use on grants, contracts and other agreements with the Federal Government, subject to the conditions in Section III.

SECTION I: INDIRECT COST RATES

RATE TYPES: FIXED FINAL PROV. (PROVISIONAL) PRED. (PREDETERMINED)

	EFFECTIVE PERIOD				
TYPE	**FROM**	**TO**	**RATE(%)**	**LOCATION**	**APPLICABLE TO**
FINAL	07/01/2014	06/30/2017	60.00	On-Campus	Research
PRED.	07/01/2017	06/30/2018	60.00	On-Campus	Research
FINAL	07/01/2014	06/30/2017	26.00	Off-Campus	Research
PRED.	07/01/2017	06/30/2018	26.00	Off-Campus	Research
FINAL	07/01/2014	06/30/2017	29.40	Mod-Off Camp	Research
PRED.	07/01/2017	06/30/2018	29.40	Mod-Off Camp	Research
FINAL	07/01/2014	06/30/2017	53.00	LDEO On-Camp	Research
PRED.	07/01/2017	06/30/2018	53.00	LDEO On-Camp	Research
FINAL	07/01/2014	06/30/2017	26.00	LDEO Off-Cam	Research

Figure 1: Excerpt from Columbia University NICRA

1. A first request generated text files of nearly 6000 unique NICRAs spanning fiscal years 1990-2007, including at least one year of information for over 3000 institutions.

2. A second request returned a spreadsheet of rate agreements for major educational institutions (about 250 major colleges and universities) covering the 1982- 2017 period. Though the first FOIA covers more institutions, and in ways is more reliable (since we know exactly what rates are being reported from the rate agreements), the second spans a longer period of time.

To create a dataset that would be useful for analyses of ICR in general, we extracted key rate information from the raw NICRA agreements (i.e. from the first FOIA) including fiscal years covered by the agreements, rate type, the specific campuses or types of research covered by the rate, institution name, and the negotiated rate itself. To create a second dataset useful for analyses focused on ICR at the NIH–the focus of much of the policy discussion regarding ICR today and historically, and of our own research–we use these data but bring in information from the second FOIA as well. For this NIH specific dataset we focused on the 1982-2007 period. A third dataset contains imputed rates for the NIH grantees for the 1982-2007 period.

2 Directory structure and files

The data and code are available at `http://www.github.com/bhavensampat/icrr`. The folders contain the raw and processed data, Python 3 code to extract data from the raw text files, and the Stata 14 code to clean and process the data. Specifically:

- `1_nicra_text/`: contains the raw text of 5849 negotiated indirect cost agreements from our first FOIA request ("FOIA 1")
- `2_nicra_clean/`:
    - `code/`: contains the Python 3 script to extract data from the FOIA 1 raw text files and the Stata 14 code to ensure all scraped data are correctly separated into their respective fields.
    - `output/`: contains the original extraction from the Python 3 script (`nicra_raw.csv` and the processed data from the Stata 14 code in Stata and CSV format
- `3_hhs/`: contains the raw and digitized PDF of rates by institution for major educational institutions from the second FOIA request ("FOIA-2"), and processed data in Stata and CSV format
- `4_nih_dataset/`
    - `code/`: contains Stata 14 code to standardize the data, restrict the data to potential NIH grantees over the 1982-2007 period, and bring in consolidated information on negotiated rates from FOIA 1 and FOIA 2
    - `output/`: contains the resulting NIH panel dataset in Stata and CSV format as well as concordances mapping the institution names from their original form (in the FOIA data) to their standardized form (to link to NIH funding databases)
- `99_byhand/`: contains CSV files with hand entered corrections and the list of "relevant" institutions for our analysis.

Below, we provide a basic overview of the processing, explain important choices made in the code, and discuss how we validated the final rate data.

3 The FOIA data

3.1 FOIA 1: Information from raw rate agreements

Our first data source is the raw text of NICRA agreements obtained via the first FOIA (FOIA 1). This request returned 5849 unique agreements, mostly for universities and hospitals. All raw agreement text files are in the /1_nicra_txt/ folder. We extracted data from these agreements using a Python 3 script (located at /2_nicra_clean/code/agreement_scrape.py). We extracted all relevant elements of each agreement to the file /2_nicra_clean/output/nicra_raw.csv. There were some idiosyncrasies in these data (inconsistent field names, missing geographic details, etc.). We corrected these using the code in /2_nicra_clean/code/nicra_clean.do. A small number of agreements were unparseable using the standard code (e.g., those where the Department of Defense was the cognizant agency and the file structure was thus different). For these agreements, we manually extracted the information and appended them to the output. Those observations that we added manually can be found at /99_byhand/byhand.xlsx.

The final output file, reflecting all agreement data from FOIA 1 after extraction and cleaning, is in /2_nicra_clean/output/nicra_cleaned.csv. There is also a Stata version (/2_nicra_clean/output/nicra_cleaned.dta.) The dataset includes the following variables:

```

. codebook, compact

Variable        Obs Unique      Mean   Min   Max  Label
-----------------------------------------------------------------------
filename      21946   5797         .     .     .  agreement identifier
institution   21946   3344         .     .     .  name of the insti...
city          21945   1221         .     .     .  city
state         21910     66         .     .     .  state
zip_code      21946   2964         .     .     .  zip
country       21279     11         .     .     .
agreement_~e  21942   2150         .     .     .  agreement final date
rate_type     21926      9         .     .     .  type of rate
rate          17243   1030  36.73444     0   120  rate
effective_~m  21926    309         .     .     .  date effective from
effective_to  17169    198         .     .     .  date effective to
special_re~k   1480    246         .     .     .  other remarks
agency        21946      3         .     .     .  cognizant agency
director      21940     76         .     .     .  director cognizan...
representa~e  21946    111         .     .     .  representative co...
telephone     21918     32         .     .     .  phone number repr...
location      21946    190         .     .     .  campus location f...
applicable    21946    296         .     .     .  function for rate
effyear       21926     28  2000.772  1983  2011  effective year of...
-----------------------------------------------------------------------

. capture log close
```

The dataset includes rate information for 3344 unique institutions. However, as Figure 2 suggests, most of the data are for the mid-1990s to early 2000s.

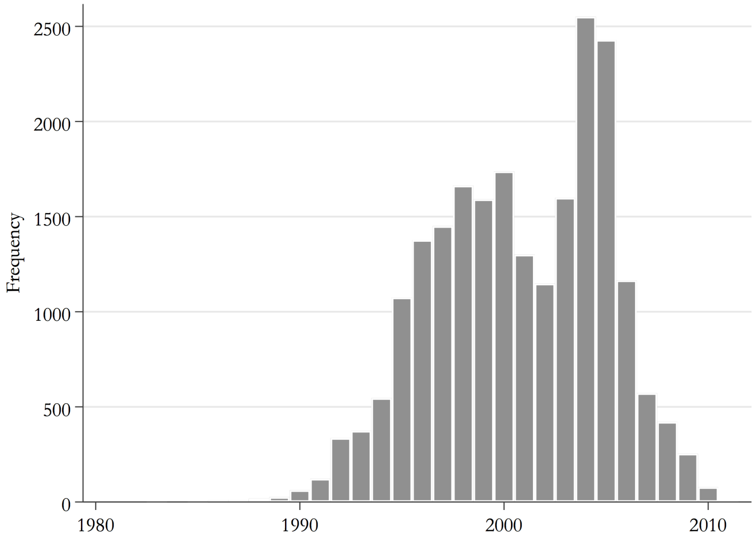


Figure 2: Distribution of agreement effective dates from `nicra_cleaned.csv`

For the 3344 institutions, the median number of fiscal years for which there is FOIA 1 rate data is 2, and the mean is 3.1. Figure 3 shows the distribution:

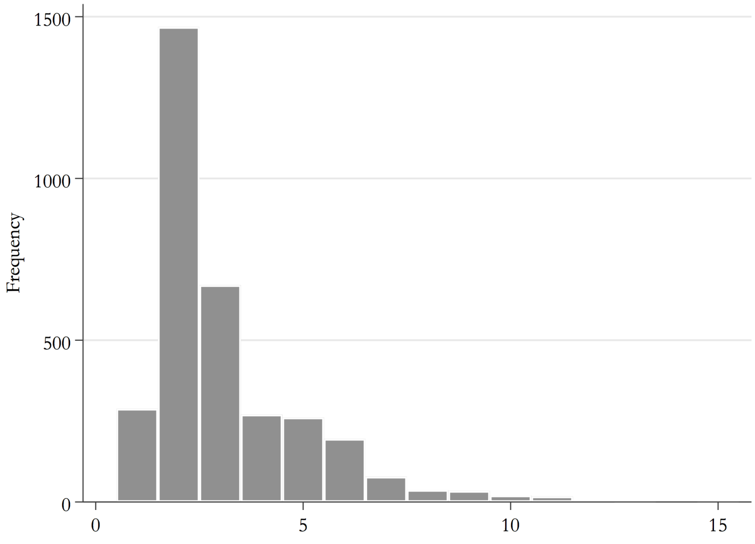
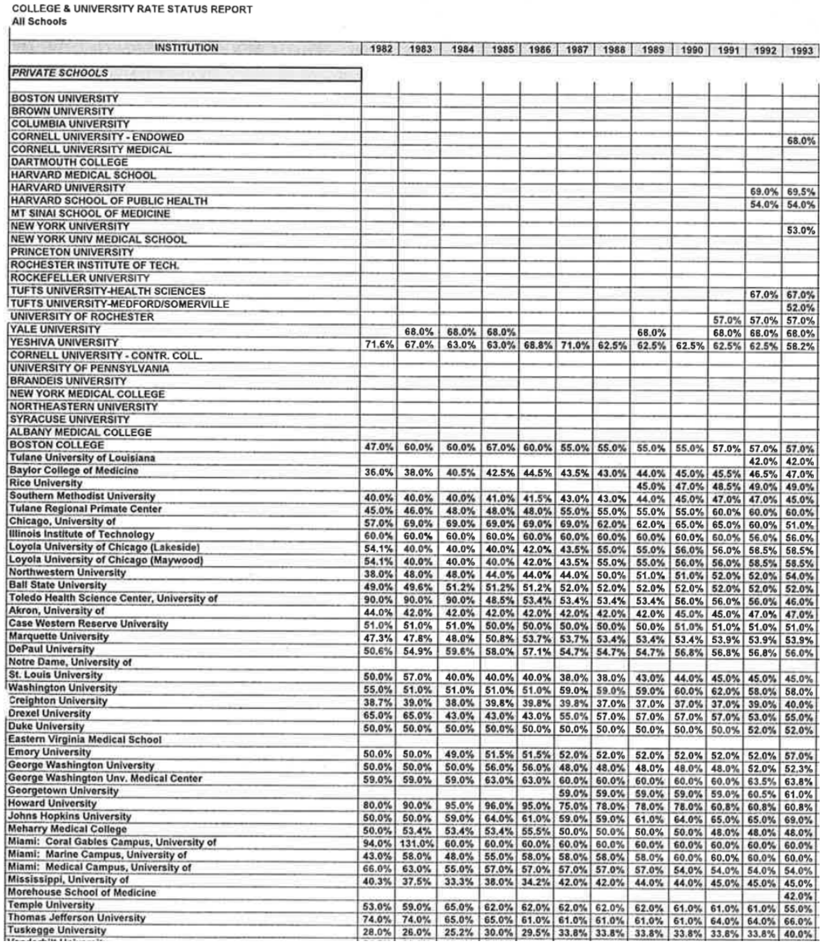


Figure 3: Histogram of number of years of data for institutions in `nicra_cleaned.dta`

The FOIA 1 rate data extracted from the raw agreements are precise: when we see an institution-rate-year we know exactly what it is (on campus or off campus rate? research rate vs. other? which campuses does it apply to), the limited time span means conducting some types of longitudinal analyses would not be possible using these data alone. Accordingly, we also provide information from the second FOIA (FOIA 2), described in more detail immediately below.

3.2 FOIA 2: Information from HHS Spreadsheet

The FOIA 2 data were obtained from a more recent FOIA to the NIH, which returned an image PDF file "College and University Rate Report" with rates for public and private colleges and universities from 1982 onward, with rate information for some institutions more consistently populated than others. Figure 3 shows an excerpt from this file:



COLLEGE & UNIVERSITY RATE STATUS REPORT
All Schools

INSTITUTION	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993
PRIVATE SCHOOLS												
BOSTON UNIVERSITY												
BROWN UNIVERSITY												
COLUMBIA UNIVERSITY												
CORNELL UNIVERSITY - ENDOWED												68.0%
CORNELL UNIVERSITY MEDICAL												
DARTMOUTH COLLEGE												
HARVARD MEDICAL SCHOOL												
HARVARD UNIVERSITY											69.0%	69.5%
HARVARD SCHOOL OF PUBLIC HEALTH											54.0%	54.0%
MT SINAI SCHOOL OF MEDICINE												
NEW YORK UNIVERSITY												53.0%
NEW YORK UNIV MEDICAL SCHOOL												
PRINCETON UNIVERSITY												
ROCHESTER INSTITUTE OF TECH.												
ROCKEFELLER UNIVERSITY												
TUFTS UNIVERSITY-HEALTH SCIENCES											67.0%	67.0%
TUFTS UNIVERSITY-MEDFORD/SOMERVILLE												52.0%
UNIVERSITY OF ROCHESTER										57.0%	57.0%	57.0%
YALE UNIVERSITY		68.0%	68.0%	68.0%				68.0%		68.0%	68.0%	68.0%
YESHIVA UNIVERSITY	71.6%	67.0%	63.0%	63.0%	68.8%	71.0%	62.5%	62.5%	62.5%	62.5%	62.5%	58.2%
CORNELL UNIVERSITY - CONTR. COLL.												
UNIVERSITY OF PENNSYLVANIA												
BRANDEIS UNIVERSITY												
NEW YORK MEDICAL COLLEGE												
NORTHEASTERN UNIVERSITY												
SYRACUSE UNIVERSITY												
ALBANY MEDICAL COLLEGE												
BOSTON COLLEGE	47.0%	60.0%	60.0%	67.0%	60.0%	55.0%	55.0%	55.0%	55.0%	57.0%	57.0%	57.0%
Tulane University of Louisiana											42.0%	42.0%
Baylor College of Medicine	36.0%	38.0%	40.5%	42.5%	44.5%	43.5%	43.0%	44.0%	45.0%	45.5%	46.5%	47.0%
Rice University								45.0%	47.0%	48.5%	49.0%	49.0%
Southern Methodist University	40.0%	40.0%	40.0%	41.0%	41.5%	43.0%	43.0%	44.0%	45.0%	47.0%	47.0%	45.0%
Tulane Regional Primate Center	45.0%	46.0%	48.0%	48.0%	48.0%	55.0%	55.0%	55.0%	55.0%	60.0%	60.0%	60.0%
Chicago, University of	57.0%	69.0%	69.0%	69.0%	69.0%	69.0%	62.0%	62.0%	65.0%	65.0%	60.0%	51.0%
Illinois Institute of Technology	60.0%	60.0%	60.0%	60.0%	60.0%	60.0%	60.0%	60.0%	60.0%	60.0%	56.0%	56.0%
Loyola University of Chicago (Lakeside)	54.1%	40.0%	40.0%	40.0%	42.0%	43.5%	55.0%	55.0%	56.0%	56.0%	58.5%	58.5%
Loyola University of Chicago (Maywood)	54.1%	40.0%	40.0%	40.0%	42.0%	43.5%	55.0%	55.0%	56.0%	56.0%	58.5%	58.5%
Northwestern University	38.0%	48.0%	48.0%	44.0%	44.0%	44.0%	50.0%	51.0%	51.0%	52.0%	52.0%	54.0%
Ball State University	49.0%	49.6%	51.2%	51.2%	51.2%	52.0%	52.0%	52.0%	52.0%	52.0%	52.0%	52.0%
Toledo Health Science Center, University of	90.0%	90.0%	90.0%	48.5%	53.4%	53.4%	53.4%	53.4%	56.0%	56.0%	56.0%	46.0%
Akron, University of	44.0%	42.0%	42.0%	42.0%	42.0%	42.0%	42.0%	42.0%	45.0%	45.0%	47.0%	47.0%
Case Western Reserve University	51.0%	51.0%	51.0%	50.0%	50.0%	50.0%	50.0%	50.0%	51.0%	51.0%	51.0%	51.0%
Marquette University	47.3%	47.8%	48.0%	50.8%	53.7%	53.7%	53.4%	53.4%	53.4%	53.9%	53.9%	53.9%
DePaul University	50.6%	54.9%	59.6%	58.0%	57.1%	54.7%	54.7%	54.7%	56.8%	56.8%	56.8%	56.0%
Notre Dame, University of												
St. Louis University	50.0%	57.0%	40.0%	40.0%	40.0%	38.0%	38.0%	43.0%	44.0%	45.0%	45.0%	45.0%
Washington University	55.0%	51.0%	51.0%	51.0%	51.0%	59.0%	59.0%	59.0%	60.0%	62.0%	58.0%	58.0%
Creighton University	38.7%	39.0%	38.0%	39.8%	39.8%	39.8%	37.0%	37.0%	37.0%	37.0%	39.0%	40.0%
Drexel University	65.0%	65.0%	43.0%	43.0%	43.0%	55.0%	57.0%	57.0%	57.0%	57.0%	53.0%	55.0%
Duke University	50.0%	50.0%	50.0%	50.0%	50.0%	50.0%	50.0%	50.0%	50.0%	50.0%	52.0%	52.0%
Eastern Virginia Medical School												
Emory University	50.0%	50.0%	49.0%	51.5%	51.5%	52.0%	52.0%	52.0%	52.0%	52.0%	52.0%	57.0%
George Washington University	50.0%	50.0%	50.0%	56.0%	56.0%	48.0%	48.0%	48.0%	48.0%	48.0%	52.0%	52.3%
George Washington Unv. Medical Center	59.0%	59.0%	59.0%	63.0%	63.0%	60.0%	60.0%	60.0%	60.0%	60.0%	63.5%	63.8%
Georgetown University						59.0%	59.0%	59.0%	59.0%	59.0%	60.5%	61.0%
Howard University	80.0%	90.0%	95.0%	96.0%	95.0%	75.0%	78.0%	78.0%	78.0%	60.8%	60.8%	60.8%
Johns Hopkins University	50.0%	50.0%	59.0%	64.0%	61.0%	59.0%	59.0%	61.0%	64.0%	65.0%	65.0%	69.0%
Meharry Medical College	50.0%	53.4%	53.4%	53.4%	55.5%	50.0%	50.0%	50.0%	50.0%	48.0%	48.0%	48.0%
Miami: Coral Gables Campus, University of	94.0%	131.0%	60.0%	60.0%	60.0%	60.0%	60.0%	60.0%	60.0%	60.0%	60.0%	60.0%
Miami: Marine Campus, University of	43.0%	58.0%	48.0%	55.0%	58.0%	58.0%	58.0%	58.0%	60.0%	60.0%	60.0%	60.0%
Miami: Medical Campus, University of	66.0%	63.0%	55.0%	57.0%	57.0%	57.0%	57.0%	57.0%	54.0%	54.0%	54.0%	54.0%
Mississippi, University of	40.3%	37.5%	33.3%	38.0%	34.2%	42.0%	42.0%	44.0%	44.0%	45.0%	45.0%	45.0%
Morehouse School of Medicine												42.0%
Temple University	53.0%	59.0%	65.0%	62.0%	62.0%	62.0%	62.0%	62.0%	61.0%	61.0%	61.0%	55.0%
Thomas Jefferson University	74.0%	74.0%	65.0%	65.0%	61.0%	61.0%	61.0%	61.0%	61.0%	64.0%	64.0%	66.0%
Tuskegge University	28.0%	26.0%	25.2%	30.0%	29.5%	33.8%	33.8%	33.8%	33.8%	33.8%	33.8%	40.0%

Figure 4: Excerpt from HHS Spreadsheet (FOIA-2)

We digitized this file, in the process correcting some minor typographical errors. We focused on cleaning rate data until 2007. The cleaned data (in wide and long form) and the original PDF file are located in the /3_hhs/ directory in CSV and Stata format. This dataset includes rate information for 250 institutions over the 1982-2007 period:

```

. codebook, compact

Variable       Obs Unique      Mean   Min   Max  Label
-----------------------------------------------------------------------
foia2_inst~n  5046    250         .     .     .
year          5046     26  1996.001  1982  2007
rate          5046    274  47.87465   8.6   131
-----------------------------------------------------------------------


. capture log close
```

The number of institutions for which there is rate information increases steadily by fiscal year:

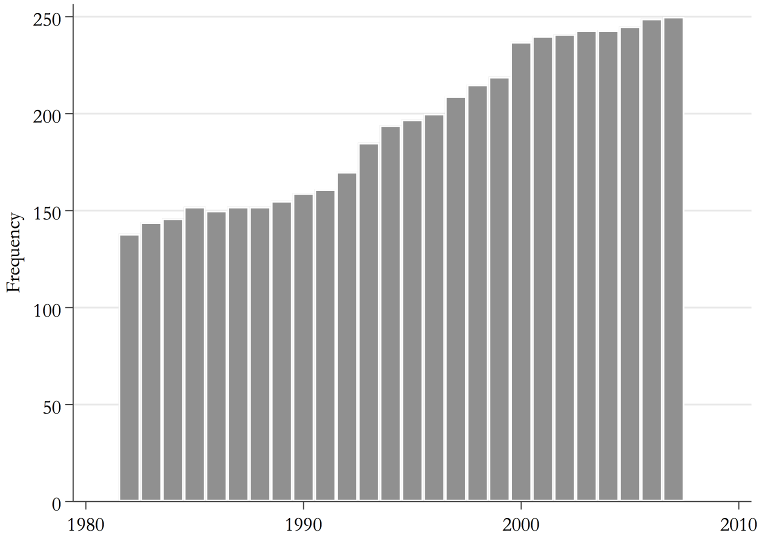


Figure 5: Distribution of agreement fiscal years from `foia2_long.dta`

Unlike FOIA 1, for FOIA 2 the majority of institutions report many years of data; Figure 6 shows a histogram of number of years of data (across the 1982-2007 period) per institution from this file:

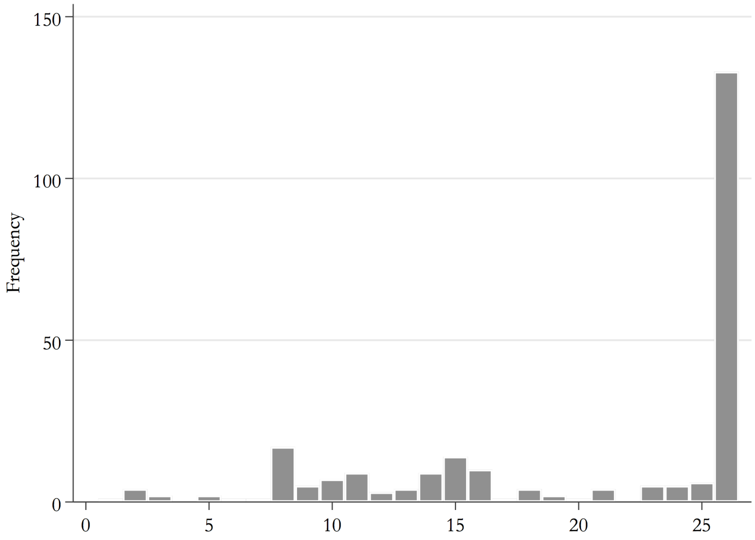


Figure 6: Histogram of number of years of data for institutions in `foia2_long.dta`

4 Creating a consolidated panel dataset for analysis of the NIH

As discussed, the two datasets discussed above have their respective advantages and drawbacks. FOIA 1 has information from actual rate agreements for a large number of institutions (including not just universities but also non-profits, hospitals, etc), but for a shorter span of time, and many institutions have only a few years of data. FOIA 2 has fewer institutions and more opaque rate data than available from the actual agreement (e.g. is the rate given on campus, off campus; a fiscal year or calendar year rate? etc.), and in some cases it is unclear what specific parts of an institution the rate in the spreadsheet reflects.

We next focused on creating a consolidated panel dataset of ICR rates for major NIH recipients over the 1982-2007 period. (We chose this time period and agency since it is the focus of much of our own analysis, but the general approach below may be useful to constructing similar panel datasets for other agencies/periods.)

4.1 Matching NIH sample to NICRA agreements (FOIA 1)

We began by isolating institutions that were "at risk" of receiving an NIH grant, and attempted to determine which of these had rate information from the two sources. To create an at risk set, we used data from the NIH's Consolidated Grant Application File (CGAF), which contains data on all NIH grants between 1938 and

2007. Using this file, we isolated a set of relevant institutions, which we define as institutions that:

1. Appear in at least three separate years from 1982 to 2007; and,
2. Received at least five separate grants over those years.

There were 736 institutions that met these criteria. These institutions account for 89 percent of total NIH funding over the 1982 to 2007 period. Though we do not include the CGAF file in this data package (since it contains non-public data) the NIH-relevant institution names are linkable to public use NIH files as well.[2]

The next step was to match to the NICRA data (`nicra_cleaned.csv`). The institution names from the NIH data did not always match those from `institution` field in the NICRA data. Beyond name variations (e.g., "UNIV" vs. "UNIVERSITY"), some NIH institutions corresponded to "subinstitutions" covered by a larger NICRA (e.g., Harvard School of Public Health and Harvard Medical School were sometimes contained in the larger Harvard University agreement). Through a mix of programming and manual cleaning we created a concordance between the standardized institutions name and any subinstitutions (located at `4_nih_dataset/output/relevant_institution_concordance.csv`). Using these standardized institution names, we merged information from `/2_nicra_clean/output/nicra_cleaned.csv)` to the CGAF grant data. The code for these steps can be found at `/4_nih_dataset/code/nih_nicra_cleaning.do`.

Of the 736 relevant institutions, there were 587 institutions for which we obtained at least one NICRA agreement over the 1982 to 2007 period. These 587 institutions account for 85 percent of total NIH funding over the 1982 to 2007 period. However, not surprisingly given our overview of the NICRA dataset above, for most institutions we had only a few years of raw NICRA data.

Most agreements contain multiple rates for different locations and activities. Since we are focused here on the NIH we isolated on campus (or on site) medical research rates for each institution. Furthermore, institutions negotiate different types of rates: provisional, final, predetermined, and fixed. Predetermined rates, generally covering a 2-4 year period, make up the most common rate type used by major research institutions contracting with the NIH. For our purposes, we include all rate types meeting the criteria above.[3]

4.2 Incorporating the HHS data (FOIA 2)

After isolating the FOIA 1 rates for the NIH-relevant institutions, we incorporated information from the HHS FOIA spreadsheet (FOIA 2) which listed rates for major

[2] Including NIH RePORTER: `https://exporter.nih.gov/`

[3] Further information on the different rate types can be found at `https://www.doi.gov/ibc/services/finance/indirect-cost-services/faqs`.

educational institutions. As we did for FOIA 1, we created a concordance mapping institution names between datasets (located at `/4_nih_dataset/output/foia2_-institution_concordance.csv`). We matched 230 of the 587 NIH "at risk" institutions to an institution in FOIA 2; the majority of those that did not match were not educational institutions and thus we would not expect them to match.[4]

To re-iterate: we consider data from the raw NICRA agreements, extracted to `nicra_cleaned.csv)` to be the most reliable source, since we know exactly which rates are being reported and for which units of an institution, and what periods of time are covered. Accordingly, we keep institutions from FOIA 2 in our final 1982-2007 NIH dataset only in cases where we can validate at least one year of data versus FOIA 1 (the raw NICRA agreements). Specifically, we incorporate rows from FOIA 1 only where:

1. There is at least one overlapping rate observation between the FOIA 1 and FOIA 2 rates
2. All overlapping rates between FOIA 1 and FOIA 2 must match perfectly

Of the 230 institutions, 170 met these validation criteria. These 170 institutions account for 96 percent of grant funding for the 736 relevant institutions. To illustrate how we consolidate the two data sources to construct the NIH panel, Figure 7 shows the sources of rate information for the top 25 institutions by total NIH funding over the 1982 to 2007 period, where:

- `.` indicates that we have no rate information available for that institution-year pair
- `1` indicates that we only have FOIA 1 (NICRA) rates for that institution-year pair
- `2` indicates that we only have FOIA 2 (HHS) rates for that institution-year pair
- `3` indicates that we have both FOIA 1 (NICRA) and FOIA 2 (HHS) rates for that institution-year pair

[4] Examples of non-matching institutions between the two sources of data include *Massachusetts General Hospital*, *San Francisco General Hospital*, and the *Oklahoma Medical Research Foundation*. Some large educational educations—such as MIT or Carnegie Mellon University—are also missing from the HHS FOIA Spreadsheet presumably because the cognizant agency in charge of negotiating their rate agreements is the Department of Defense (DoD), rather than HHS.

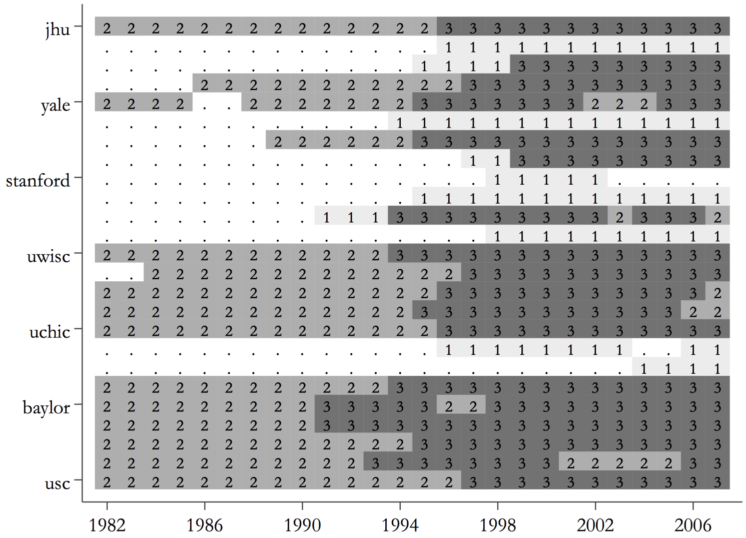


Figure 7: Source of rate data (FOIA 1, FOIA2, both) for top 25 institutions (by total NIH funding 1982-2007) listed in /4_nih_dataset/output/nicra_panel.csv, by year

Here each row represents an institution. Institutions are listed in descending order of total funding. For the top institution, Johns Hopkins, about half of the rate date are from FOIA 2, and the other half listed in both sources. In some cases, we only have data from FOIA 1. Note, however, that we only include FOIA 2 information where there is also FOIA 1 information that matches exactly (i.e. there are no rows with a 2 alone).

4.3 Consolidated NIH Panel Dataset: Original Rate Data

The resulting dataset /4_nih_dataset/output/nicra_panel.csv and /4_nih_dataset/output/nicra_panel.dta includes the following variables:

```

. codebook, compact

Variable         Obs Unique      Mean   Min   Max  Label
---------------------------------------------------------------
institution   229632    736         .     .     .  standardized ins...
agreement_~e   53030    994         .     .     .  date of underlyi...
month         229632     12       6.5     1    12  month code (1 = ...
year          229632     26    1994.5  1982  2007  calendar year
rate_type      53030      6         .     .     .  rate type from I...
rate1          53030    352  52.25226     0   120  rate from ICR ag...
rate2          41044    221  48.06236   8.6   131  rate from NIH FO...
```

```
rate_n         74852    412  50.78496     0   131  composite rate o...
location       99518      4         .     .     .  location informa...
applicable     99518      3         .     .     .  applicable infor...
city           53030    311         .     .     .  institution city...
state          53030     52         .     .     .  institution stat...
zip_code       53030    656         .     .     .  institution zip ...
agency         53030      2         .     .     .  negotiating agen...
director       53030     52         .     .     .  agency director ...
representa~e   53030     80         .     .     .  agency represent...
telephone      53006     26         .     .     .  agency/represent...
filename       53030   1404         .     .     .  filename (unique...
special_re~k    2302     50         .     .     .  other informatio...
country        51889      1         .     .     .  institution coun...
original_f~m   99518    141         .     .     .  original date of...
original_to    99518    138         .     .     .  original date of...
----------------------------------------------------------------------

. capture log close
```

There are 736 unique institutions, and a potential of 26 years of data for each. Since data are the monthly level, this means there are $736 \times 26 \times 12 = 229,632$ potential observations. For each institution-month observation we include data from FOIA 1 (`rate1`), a rate from FOIA 2 (`rate2` *only provided if it has been "validated", i.e. there is at least one other year of overlap with FOIA 1 and any overlapping rates match perfectly*), and a final rate (`rate_n`) which is the FOIA 1 rate where it exists, or a "validated" FOIA 2 rate. If we focus on the June rate for each institution (month==6) as the rate for that year, we calculate that 80 percent of the 736 institutions have at least 1 year with rate data, and 40 percent have at least 10 years of data.

Figure 8 shows the distribution across institutions of the number of years where `rate_n` is populated (again based arbitrarily on the June rate for that year). In addition to the overall distribution across the 736 institutions, we also examine this by quintile of total funding over the entire period:

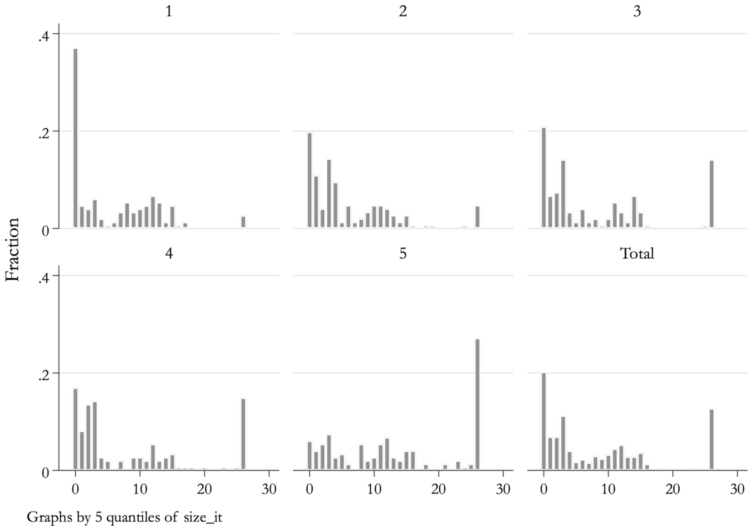


Figure 8: Distribution across institutions of the number of years of non-missing rate data in `nicra_panel.csv`, overall and by quintile of total 1982-2007 funding (5=top)

Overall, as seen in the final panel) there is a full panel of data (26 years) for only 13 percent of the institutions (94/736), and 20 percent (149/736) have no rate data. However, the data coverage is much better for top funding quintiles. For example, for the top quintile there is at least 10 years of data for 60 percent of the institutions.

4.4 Consolidated NIH Panel Dataset: Imputed Data

To improve the panel coverage of the dataset, we also experimented with imputing missing rates. Specifically, we used grant level data from the CGAF on actual direct and indirect costs paid per grant to try to predict the negotiated rate `rate_n`. The imputed rates are in the final output file we provide (`/4_nih_dataset/output/imputed_rates.csv`).

Specifically, using CGAF data on actual direct and indirect cost dollars for each grant awarded between 1982 and 2007, we calculated the actual indirect cost rate at the grant level. Previous analyses of this sort have emphasized that actual rates paid at the grant level could differ from negotiated rates because certain grants (e.g., training grants) cap or disallow indirect costs, some policies restrict indirect costs paid on certain types of expenses, and various subcontract rules may drive a wedge between actual and negotiated rates.[1]

Consider for example fiscal year 2000. Among the 736 “at risk” institutions, we had negotiated rate information (`rate_n`) for 360. Using CGAF data on all grants awarded to these institutions in that fiscal year, we also calculated the modal and median actual rates paid. Figure 9 shows a scatterplot of `rate_n` versus the median and modal actual rates:

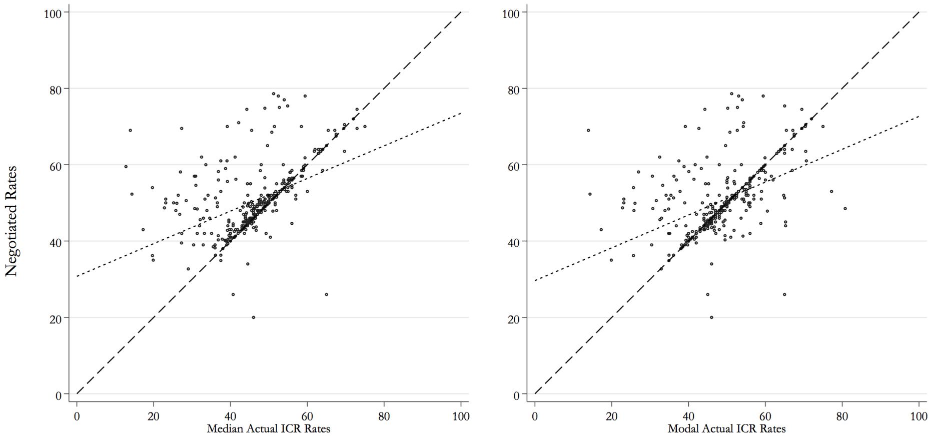


Figure 9: Negotiated ICR rates vs. median and modal actual rates paid on grants, FY 2000)

The dashed 45° line indicated exact matches between negotiated and imputes rates. For a non-trivial share of the observations there is exact match: 35 percent match exactly using the modal actual rate, and 27 percent using the median actual rate. The solid line is a fitted regression line, and indicates a positive relationship between the actual and negotiated rates. It is also apparent that for most institutions the actual rates tend to be less than the negotiated rates, consistent with the research cited above.

We inspected the non-matches by hand, and reviewed grant level ICR guidelines to understand the potential reasons that grant level rates would deviate from negotiated rates. We conjectured that grants meeting the following conditions were more likely to reflect the negotiated rate:

- R01, R23, R29, and R37 grants
- Grants with simple structures: those without without amendments, supplements, and subprojects
- Grants awarded through the larger NIH institutes and centers of the NIH
- Grants where direct cost awards were less than $500,000.

In general, we expect these criteria to help focus on "no frills" grants that are not subject to caps or other restrictions (and therefore more likely to reflect the negotiated rate).

Overall, in cases where we have both negotiated rates (`rate_n`) and can predict a rate based on modal grants of the type above, the rates match 45 percent of the time using the modal rate and 40 percent of the time based on the median, an improvement on the raw mode and median (across all grants) presented above. Both the match rates and correlations are much higher for more heavily funded institutions—again measured by quintile of funding—as Figures 11 and 12 indicate. For the mode, the match rates across-institution years are 21 percent for bottom quintile, 38 percent for second, 46 percent for third, 55 percent for fourth, and 64 percent for the top quintile of institutions by funding. (The match shares are very similar for medians.) Visually, we see that the fitted regression line approaches the 45° line for the top quintiles.

As a practical matter, the imputed rates may be useful for some analyses (especially those focused on large institutions), but the imperfect matching rates indicate they should be used with caution. These rates (based on the mode) are provided in the file `/4_nih_dataset/output/imputation_rates.csv` and `/4_nih_dataset/output/imputation_rates.dta`. In this file, 62 percent of the potential institution-year observations between 1982 and 2007 are populated with a rate, as compared to 32 percent based on the real (non-imputed) data in `4_nih_dataset/output/nicra_panel`. We are currently experimenting with machine learning approaches to improve the match rate as well, and will update the imputed rate file if these perform substantially better than brute force imputation based on modal and median actual rates.

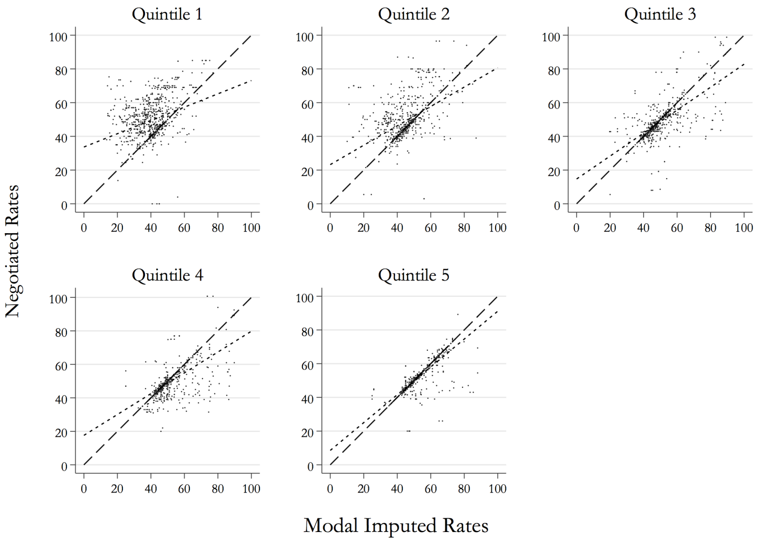


Figure 10: Negotiated rates versus imputed rate (using modal rate for “no-frills” grants) by quintile of NIH funding in a fiscal year (5=top)

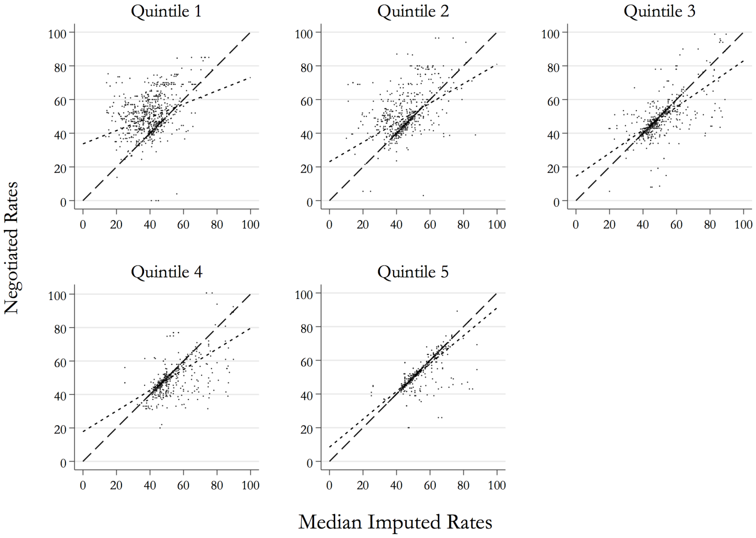


Figure 11: Negotiated rates versus imputed rate (using median rate for "no-frills" grants) by quintile of NIH funding in a fiscal year (5=top)

References

[1] Heidi Ledford. Indirect costs: Keeping the lights on. *Nature News*, 515(7527): 326, 2014.