WELLESLEY

Deciphering the Innovation Pipeline of Physician-Scientists

Ashley You, Fall ‘24, Data Science Major Capstone

Background and Research Questions

Physician-scientists, who are proficient in both medicine and research, have historically been crucial members of the biomedical workforce, facilitating the translation of research findings from clinical settings to laboratories and vice versa. Despite their significance, there remains a lack of research to systematically assess their long-term impact and the variables that shape their training and career trajectories.

The objective of this project is to uncover the characteristics that define a physician-scientist as an innovator and to elucidate the steps involved in translating their research from conception to commercialization.

This poster will illustrate:

- **Patterns observed in outputs such as patents, grants, clinical trials, and publications from ASCI 1995 and 2010 cohort.**
- **How grants play a factor in influencing the creation of patents, clinical trials, and publications from ASCI 1995 and 2010 cohort.**

Data Collection

Being limited to open-sourced data, I extracted data from:

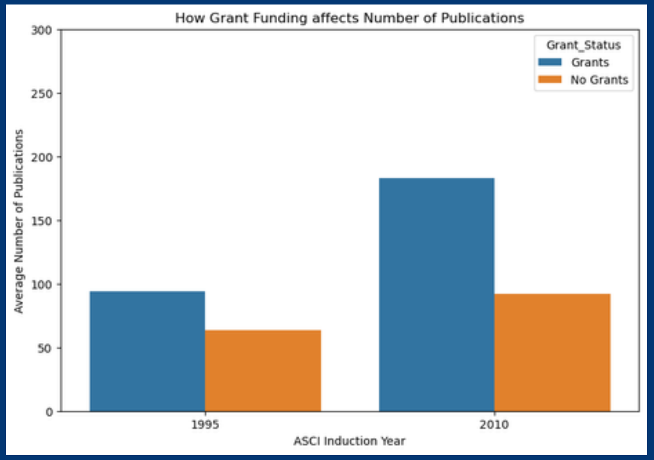
- The American Society for Clinical Investigation (ASCI) Website
    - Names, institutions, medical specialties
- OpenAlex Database
    - Publications
- DimensionsAI Database
    - Grants, Patents, and Clinical Trials

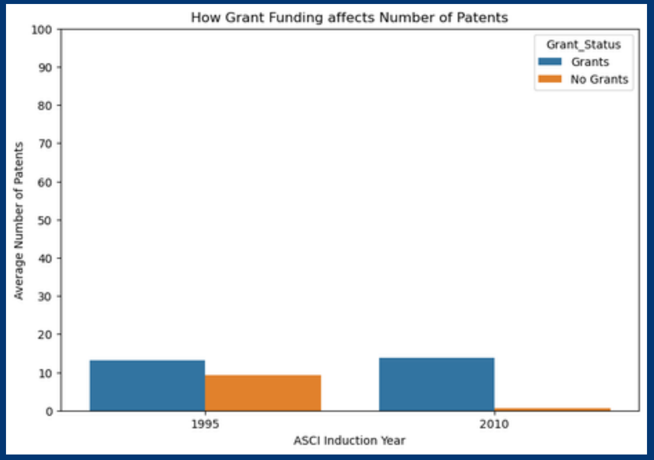
1. Identify Physician-Scientists
    a. I chose to web scrape the ASCI Directory to collect a full list of physician-scientist names that date from 1958 to 2023
        i. This poster only contains data from names inducted into ASCI in 1995 and 2010
2. Query Databases
    a. I obtained an institutional key to access both OpenAlex and DimensionsAI which are aggregator databases
    b. Both databases use author IDs when querying by name. So, I had to manually disambiguate the IDs that could be associated with a single-name search.
    c. After isolating each physician-scientist name ID(s) in each database, I was able to query for their respective works.
3. Changing the form of the data file
    a. Both databases output data queries in a JSON file, which does not make data analysis easy.
    b. I converted the JSON files into CSV files by isolating universal keys and setting them as column names

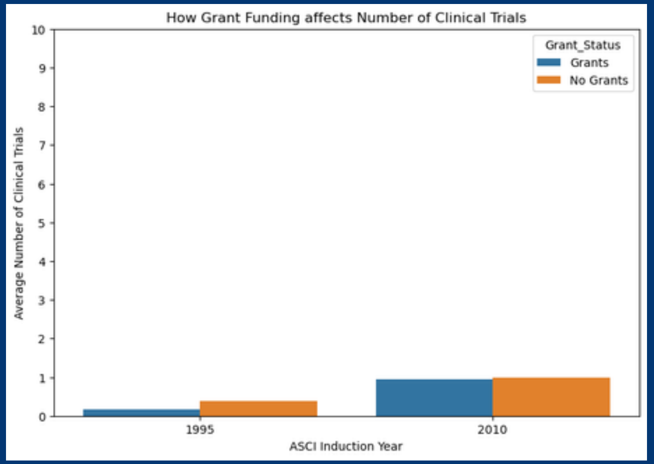
Data Cleaning

The data from both databases need much-needed cleaning.

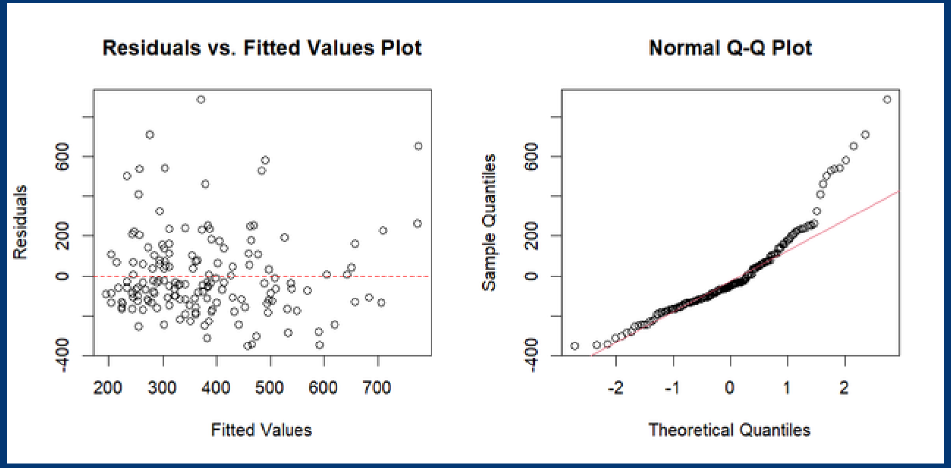
- Removed duplicated data based on titles and IDs
- Titles and variables had ASCII code that needed translation to English
- Failed queries had to be manually added to the CSV files
- created new column names to label an indiviual’s ASCI cohort year







Data Modeling: Multiple Linear Regression



The multiple linear regression model utilizes all data from the 1995 cohort and 2010 cohort. It models to see if the number of clinical trials, patents, and grants influence a physician-scientist’s number of publications.

- The model showed a positive relationship between the number of clinical trials and number of grants affecting the output of publication
- But, the plots above show violations of equal variance assumption and some departure from normality.

Pattern Observations:

The data on the left demonstrates the average production of publications, clinical trials, and grants, extracted from a decade following each ASCI induction year for an equitable analysis.

- The 1995 induction considers data from the period 1995-2010, while the 2010 induction covers data from 2010-2020. .

The sequence of output from highest to lowest is: publications, followed by patents, and then grants

Conclusion:

Since the multiple regression model violates a homoscedastic situation, the validity of the positive trend may not hold. But, our bar graphs do support the pattern of grants affecting publication output. However, the sample size did not show many occurrences of Patents and clinical trials which is a significant factor in output comparison and variance inaccuracies.

Data Ethics and Limitations

These variables do not encapsulate all the observations that could relate to and influence the support work of a physician-scientist. Given more time, I hope to disambiguate more physician-scientist names and identify more variables in order to better represent the innovation pathway of this workforce.