FACTORS ASSOCIATED WITH IMPACTFUL SCIENTIFIC PUBLICATIONS IN NIH-FUNDED HEART DISEASE RESEARCH

A Thesis

Presented to the

Faculty of

San Diego State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

in

Computer Science

by

Diana Rozenshteyn

Summer 2023

SAN DIEGO STATE UNIVERSITY

The Undersigned Faculty Committee Approves the

Thesis of Diana Rozenshteyn:

Factors Associated with

Impactful Scientific Publications

in NIH-Funded Heart Disease Research

Hajar Homayouni, Chair
Department of Computer Science

Roger Whitney
Department of Computer Science

Juanjuan Fan
Department of Mathematics and Statistics

Approval Date

DEDICATION

I would like to dedicate this work to my family, whose unwavering support throughout the six years of my career-changing degree has been invaluable.

We've tended to forget that no computer will ever ask a new question.

– Grace Hopper

ABSTRACT OF THE THESIS

Factors Associated with
Impactful Scientific Publications
in NIH-Funded Heart Disease Research
by
Diana Rozenshteyn
Master of Science in Computer Science
San Diego State University, 2023

In this study, we investigated factors associated with impactful scientific publications funded by the National Institute of Health (NIH) in the field of cardiovascular diseases (CVD). We analyzed a database of NIH-funded heart disease research publications from 2002 to 2020 to uncover key factors contributing to successful outcomes in this field. The study found that funding provided by the NIH positively correlated with the number of publications, and spending cuts in scientific research by the US Congress are associated with research productivity. Our exploratory data analysis revealed the concentration of heart disease research articles in a small number of journals and institutions. We observed gender disparities, with a higher representation of male first authors in top journals and institutions. We demonstrated that male-authored publications received more citations and had higher NIH Percentile scores compared to female-authored publications, indicating a persistent gender gap in publication and visibility. The study also employed predictive modeling, where regression models initially performed poorly, but the XGBoosting model demonstrated the best predictive power among classifiers in estimating the success of cardiovascular research. We defined the success of a publication as the number of people who benefit from the published research and we employed the NIH Percentile as an approximation of that measure. Journal Rank emerged as the most influential feature in predicting research success. Overall, this study provides important insights into the factors influencing impactful publications in the field of heart disease research, including funding, journal selection, institutional affiliations, gender disparities, and the potential for predicting research success.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PAGE

GLOSSARY

AD Author Institution Affiliation.

API Application Programming Interface.

AUC Area Under the Curve.

CVD cardiovascular diseases.

DALYs disability-adjusted life years.

DP Date of Publication.

FAU Full Author.

GBD Global Burden of Disease.

JT Journal Title.

MeSH Medical Subject Headings.

ML Machine Learning.

NHLBI National Heart, Lung, and Blood Institute.

NIH National Institute of Health.

NIH Percentile NIH Percentile.

NLM National Library of Medicine.

NSF The U.S. National Science Foundation.

OPA Office of Portfolio Analysis.

PMID PubMed Unique Identifier.

RCR Relative Citation Ratio.

ROC Receiver Operator Characteristic.

ROR Research Organization Registry.

SEH science, engineering, and health.

SJR Scientific Journal Ranking.

SMOTE Synthetic Minority Oversampling Technique.

SSN Social Security Number.

ACKNOWLEDGMENTS

I express my heartfelt gratitude to Pr. Homayouni for her invaluable insights, constructive feedback, and unwavering support during the process of working on my thesis. Her guidance has been instrumental in shaping the quality and direction of my research, and I am truly thankful for her contributions.

Furthermore, I would like to extend my sincere thanks to my thesis committee members, Pr. Whitney and Pr. Fan. I am truly grateful for their support.

Lastly, I want to convey a heartfelt appreciation to my family, who believed in my abilities during this midlife journey and provided unwavering support through all the challenges and triumphs. Your encouragement and cheers have been instrumental in my success, and I am deeply grateful for your belief in me.

CHAPTER 1
INTRODUCTION

This study examines the factors associated with impactful scientific publications funded by the National Institute of Health (NIH) in the field of cardiovascular diseases (CVD), the leading medical condition in the US and worldwide. CVD is characterized by high disease burden and mortality rates, as reported in the Global Burden of Disease (GBD) study [1], [2]. The burden of disease is quantified using disability-adjusted life years (DALYs) [3], a more comprehensive measure than mortality rates for assessing the societal and individual impact of diseases [4].

Considering CVD is the primary cause of death globally, research focused on CVD and heart disease, holds significant importance in advancing scientific knowledge and enhancing treatment and disease outcomes. However, the field of CVD research is currently confronted with inadequate funding and a shortage of researchers, particularly women, who are leaving the field [5], [6].

To analyze these issues, we have collected a database of NIH-funded heart disease research publications and related information spanning from 2002 to 2020. Using this data, our project aims to answer the following questions:

How has the total number of publications in heart disease research changed over the years, and is there a correlation between the amount of funding allocated for research and the number of publications?

Researchers face intense competition for funding, publication acceptance in prestigious journals, and positions at renowned universities and research centers. This competition is further intensified by the surplus of individuals with PhDs relative to the limited number of available research positions [7], [8]. Given these circumstances, securing funding from the NIH is of the utmost importance for researchers. However, the funding rates for a common type of NIH grant, R0-1, are relatively low. Over the 18 years examined in this study, R0-1 grant funding rates steadily declined from 34% in 2002 to 29% in 2020 [9]. Additionally, the number of applicants competing for research funding in 2020 was 1.7 times higher than in 2002. Furthermore, R0-1 grant funding rates are even lower for new investigators compared to established ones, with rates of 23% vs. 32% in 2002 and 22% vs. 30% in 2020. This project uses exploratory data analysis and linear correlation to evaluate changes in the number of publications over time and investigate whether there is a correlation between funding and the number of publications in the field of heart disease research.

What are the top publishing journals in the field of heart disease research? Are journal ranking scores influential in publishing a high number of papers? Is there a gender gap in terms of publication opportunities in these top publishing journals?

Despite efforts to move away from performance metrics like publication count and journal impact factors in decision-making processes related to funding, hiring, and publication acceptance, these metrics still hold importance in academia [10], [11], [12]. The number of scientific publications and the prestige of the journals in which they are published can significantly impact a researcher's career, including their ability to obtain funding, secure employment, or attain tenure [10]. This project aims to analyze the most frequently publishing journals in heart disease research and explore potential gender disparities in publication opportunities. Moreover, we will investigate the importance of journal ranking scores for a successful publication.

Which research institutions are the top publishers in the field of heart disease research? Is there a gender gap in publication opportunities at these institutions?

Research institutions benefit from accomplished researchers with a successful publication track record, which enhances their funding and reputation [13]. For instance, the top twenty research universities, in THE World University Rankings 2020, had a high number of publications in top-rated journals, which received double the citations compared to similar publications, based on the field-weighted citation impact score [14]. This project aims to analyze which institutions have the highest number of publications in heart disease research and investigate potential gender disparities in publication opportunities at these institutions.

Is there a difference in the number of citations and NIH Percentile between publications authored by females and males?

Gender disparity is evident in the way female and male authors are cited, as indicated by a cross-sectional study of high-impact journals in academic medicine which states that "articles written by women as both primary and senior authors had approximately half the number of citations as those authored by men as both primary and senior authors" [15]. This study will examine the difference in the number of citations for female and male-authored publications in the heart disease research field.

Is there a gender gap in heart disease research and how did it evolve over the years?

Women in STEM professions face additional challenges in a competitive scientific field. According to The U.S. National Science Foundation, in 2019, only 39% of women with doctorates in science, engineering, and health (SEH) were employed in academia, and women held only 33% of full-time senior faculty positions [16]. Furthermore, a lower percentage of female first and last authors are getting to publish their research in journals with a high impact factor [17]. In particular, there is still gender disparity in publications by female authors holding senior positions in the field of academic medicine [18]. This project aims to study the gender gap in publication opportunities in the field of heart disease research. We will investigate potential gender disparities by analyzing the journals where publications appear and the research institutions the first authors are affiliated with.

Is it possible to predict the success of a paper, as measured by NIH Percentile, based only on the information available at the time of publication and what are the key factors associated with successful publication in heart disease research?

Successful research projects in the cardiovascular domain can benefit a large population in society. Identifying institutional and demographic contributors to successful research outcomes remains an important task, especially in interventions to reduce biases. This project aims at identifying key factors in successful research outcomes in heart disease research. We use interpretable machine learning models for research impact estimation while explaining major contributors to productive outcomes, which will be measured based on the number of people who benefit from the published research.

CHAPTER 2
Data Preparation

This chapter focuses on the crucial steps of data collection and cleaning, which are essential for conducting successful exploratory data analysis and machine learning modeling. Data collection involves gathering relevant information from various sources and data cleaning involves removing inconsistencies, errors, and missing information within the collected data, ensuring data accuracy and reliability for analysis. We will provide an overview of the data collection process and discuss various techniques and methodologies employed in data cleaning to ensure the integrity and quality of the data.

2.1 Data Collection

We used PubMed's advanced search tool [19] to construct datasets of publications on cardiovascular disease. PubMed is a free research resource made available by the National Library of Medicine (NLM) to search biomedical literature. Two limiting factors, the availability of the first author's full name and the years required for citations to occur, were used to choose the time period for data gathering. PubMed records started to include full author's names, Full Author (FAU), for articles starting in 2002 [20]. In addition, three years is the minimum number of years recommended for citation and publication impact analysis [21]. Furthermore, 2020 is the last year for which the Scientific Journal Ranking (SJR) information needed for data analysis is available [22]. Therefore, we searched PubMed for records from 2002 to 2020. There are a total of 18 datasets, one for each year.

NIH funding in the study is represented by National Heart, Lung, and Blood Institute (NHLBI) grant funding. We used the NHLBI grant ([GR]), the Date of Publication (DP) keywords in queries, along with the combination of keywords that are based on the cardiovascular disease-related conditions listed on the GBD data tool site [23]. Those keywords are *cardiovascular*, *ischemic*, and *heart*.

Example of PubMed query for the year 2020: "cardiovascular OR ischemic OR heart AND NHLBI[GR] AND 2020[DP]"

To get the journal name, article's first author institution affiliation, and country for further parsing, we saved the advanced search PubMed queries by choosing the abstract format option in the display options menu and the PubMed format in the save citations to the file menu [19]. To acquire the list of PMIDs (PubMed Unique Identifier) for each publication needed for further citation information, we saved data collected via advanced search PubMed queries by choosing the abstract format option in the display options menu and the PMID format in the save citations to file menu [19].

To acquire the citation-related information and full unabbreviated names of the authors when available, we uploaded the PMIDs dataset for each year to the ICite web tool [24]. We saved the resulting data analysis as csv files. ICite is run by the Office of Portfolio Analysis (OPA). The OPA is a division of the NIH that is responsible for the data-driven evaluation of research to help the NIH decide what current or new research areas will have a greater benefit for science and human health [25]. ICite provides available information on the author’s full first name, total citations, citations per year, a field- and time-adjusted citation measure of scientific influence called Relative Citation Ratio (RCR), and NIH Percentile. “RCR represents the field- and time-normalized citation rate, and is benchmarked to 1.0 for a typical (median) NIH paper in the corresponding year of publication” [24]. NIH Percentile is the rank of any individual paper’s RCR score relative to all other NIH publications [24]. A higher NIH Percentile is indicative of a more impactful paper with a higher citation rate.

The PubMed format datasets cannot be saved in a CSV format and therefore had to be parsed to extract Journal Title (JT), first Author Institution Affiliation (AD), and country. We wrote a parsing script in Python 3.10.1 for these purposes [26]. First author affiliation was determined by making an Application Programming Interface (API) request to the Research Organization Registry (ROR) API [27]. ROR affiliation matching allowed us to find research organizations mentioned in the full affiliation strings from the PubMed format datasets which are then provided in the API call [28]. The results of the API call are returned in the JSON format [29]. Here is an example of the API call used:

```
response = requests.get(’https://api.ror.org/organizations?affiliation=’
+ urllib.parse.quote(institution_info)).json()
```

We parsed journal titles and countries from the PubMed format datasets using PubMed Data Element (Field) Descriptions included in this type of file format [20]. We processed the dataset for each year separately.

We processed the parsed PubMed format datasets for each year queried in a JupyterLab [30] and merged them on PMIDs with ICite citation datasets. No records were lost at this step and the total number of records was 121,421 for all years combined (Table 2.1-Column “At Query Time”).

We merged the resulting datasets for each year based on journal names with the SJR dataset. We downloaded the SJR dataset for the last available year, 2020, from the

Table 2.1. Number of Records at Various Data Processing Steps by Year

Year	At Query Time	After Wrong Year Filtering	% of Records Removed After Wrong Year Filtering	After Journal Rank Merge	Number of Unknown Journals After Journal Rank Merge	% of Unknown Journals After Journal Rank Merge
2002	3,793	3,625	4.43	3,625	318	8.90
2003	4,046	3,772	6.77	3,772	338	9.12
2004	4,344	3,952	9.02	3,952	337	9.16
2005	4,741	4,172	12.00	4,172	389	10.21
2006	5,013	4,316	13.90	4,316	464	12.00
2007	5,367	4,675	12.89	4,675	480	10.91
2008	5,797	5,076	12.44	5,076	590	12.78
2009	6,025	5,155	14.44	5,155	630	13.79
2010	6,613	5,680	14.11	5,680	652	12.90
2011	7,025	5,940	15.44	5,940	669	12.84
2012	7,228	6,196	14.28	6,196	668	12.57
2013	7,568	6,451	14.76	6,451	653	11.44
2014	7,924	6,799	14.20	6,799	667	11.49
2015	7,952	6,832	14.08	6,832	615	10.07
2016	7,606	6,638	12.73	6,638	520	8.84
2017	7,285	6,323	13.20	6,323	474	8.54
2018	7,451	6,474	13.11	6,474	462	8.31
2019	7,545	6,633	12.09	6,633	500	8.44
2020	8,098	6,952	14.15	6,952	545	7.84
2002-2020	121,421	105,661	13.43	105,661	9,971	9.44

SCImago Journal & Country Rank website [22]. The SCImago Journal & Country Rank database ranking used in this study is based on the SJR indicator. The SJR indicator was developed from the information contained in the Scopus® database [31] and is a measure of a journal's scientific impact. SJR is calculated by dividing weighted citations for any given year by weighted citations in that journal for the previous 3 years. The citation weight is calculated based on the journal's prestige in the particular scientific field and the scientific closeness of the cited journal to the journal that cites the article [32].

We used the Gender−API web service to estimate the gender of first authors [33]. The first author's name and country were used for this estimation. It is important to note that gender estimation tools such as Gender−API determine gender in a binary form and are unable to take into consideration the gender identity of individuals [34]. This might lead to misidentifying a person's gender. Therefore, in this study, first name serves as a proxy for gender to estimate the first author's gender identification.

In several studies that compared performance by various gender detection web tools, the Gender−API service had the best performance among the services compared based on errorCoded error [35], [36]. ErrorCoded is the proportion of misclassifications and nonclassifications over the total number of assignments. Additionally, gender detection performance is generally poor with regard to correctly identifying Asian names compared to European names across gender-identifying apps [37]. Santamaría and Mihaljević showed that Gender−API performance, evaluated as a measure of errorCoded, produced 18% inaccuracies for Asian names compared to 3% for European names [35]. NamSor [38] app performed even worse producing 35% inaccuracies for Asian names compared to just 3% for European. In a study by Sebo, Gender−API performed better in identifying Asian names compared to the Wiki−Gendersort app [39], with errorCoded being 65% for Gender−API versus 90% for Wiki−Gendersort, but performed worse than the NamSor app, which had an errorCoded value of 53% [37].

Researchers have also used the US Social Security Number (SSN) dataset for gender determination. For example, Hu et al. [40] used a predictive model based on the one-hot encoding of the first name characters to estimate gender. The model was trained using the US SSN dataset. Mehran et al. [41] also used US SSN dataset to estimate the gender of the first authors in Cardiology Randomized Clinical Trials to quantify the gender gap. However, since the US SSN dataset does not cover international names, it was not considered for this study.

2.2 Data Cleaning

We continued to process the files for each year from Section 2.1 in a JupyterLab [30] with the help of the Python language and Pandas library functions [26], [42]. Due to the different formatting of journal names in PubMed format datasets and the SCImago Journal and Country Ranking dataset, 9.44% of entries on average remained unmatched even after the data cleaning effort taken to achieve the same journal name formatting in both datasets (Table 2.1-Column "% of Unknown Journals After Journal Rank Merge"). Since the journal ranking number could not be assigned, we marked unmatched records as "unknown" in the resulting dataset for each year.

Next, we filtered the resulting dataset for each year to only include records for the year listed in the PubMed original query. Despite the fact that each PubMed query specifically searched for only one year of publications, the resulting datasets downloaded from PubMed contained years that were not in the query. By filtering out the wrong years of publications 13.43% of entries on average were removed (Table 2.1-Column “After Wrong Year Filtering”).

We merged all datasets for 2002-2020, resulting in 105,661 total records (Table 2.2, step 1). Next, we replaced the first name with an “unknown” label in 3,760 records for which the first author’s full first name was not available, (Table 2.3-Column “Number of Records that have an Unknown Feature”). This step was done to identify records for which the first author’s gender identification was not possible. 0.44% of the records were lost during the first author’s gender identification process conducted via the Gender−API service website (Table 2.2, step 2). No explanation as to why those records were not recovered after processing was provided by the Gender−API service. In addition, we marked 2,614 records with a gender identification accuracy of less than 60%, as determined by Gender−API, as “unknown” for the gender category (Table 2.3-Column “Number of Records that have an Unknown Feature”). We choose the threshold of 60% for gender identification accuracy based on the previous studies that explored gender trends in research publications [43], [44], [17].

Table 2.2. Total Number of Records at Various Data Processing Steps, 2002-2020

Steps	Data Processing Steps	Number of Records	% of Records Removed
1	Merging records for 2002-2020	105,661	N/A
2	Gender API first author’s gender identification	105,195	0.44
3	Removing records that are not articles	86,904	17.39
	Total Records removed	18,757	N/A
	Total Records after all data processing steps	86,904	17.75

In addition, we removed records that were not articles resulting in reducing the number of records by 17.39% (Table 2.2, step 3). Non-article publications include various categories, such as reviews, bibliographies, technical reports, and clinical conference reports [45]. This step also resulted in reducing the number of records for which the journal name was not determined from 9,971 (Table 2.1) to 6,912 (Table 2.3).

Table 2.3. Number of Records with Unknown Features, 2002-2020

Feature	Reason for Unknown Feature	Number of Records that have an Unknown Feature
First author's full first name		3,760
Country		61
First author's gender	no full first name available	3,760
First author's gender	identification accuracy is less then 60%	2,614
First author's institution affiliation		129
Journal name		6,912
Total number of records: 86,904		

We marked 61 records for which the country of the first author was not possible to determine as "unknown" in the country category (Table 2.3).

Finally, in the institution category, we also marked as "unknown" 129 records for which the first author's institution affiliation was incorrectly assigned or not found during the ROR API call (Table 2.3).

In summary, as a result of data collection and cleaning processes (Table 2.1 - 2.3), the number of records was reduced by 28.42% from 121,421 to 86,904.

CHAPTER 3

Exploratory Data Analysis

Exploratory data analysis plays a crucial role in understanding the characteristics of the data, identifying patterns, and extracting meaningful insights. We use various statistical and visualization techniques to explore the dataset and gain a deeper understanding of its structure, distributions, relationships, and potential outliers. In particular, we look at the role funding plays in research outcomes in the heart disease research field. Furthermore, we investigate potential relationships between gender and other variables of interest, such as funding, institutional affiliations, journals, journal rankings, and citation output. By examining these associations, we can identify any potential biases or disparities that may impact research outcomes.

3.1 Change in Heart Disease Research Publications over the Years

The number of heart disease research publications grew over the 18 year period by 43.6% from 3,164 in 2002 to 5,587 in 2020 (Figure 3.1). However, the number of publications increased 40.5% by, and including, 2013 and only increased by 4.7% between 2013 and 2020.

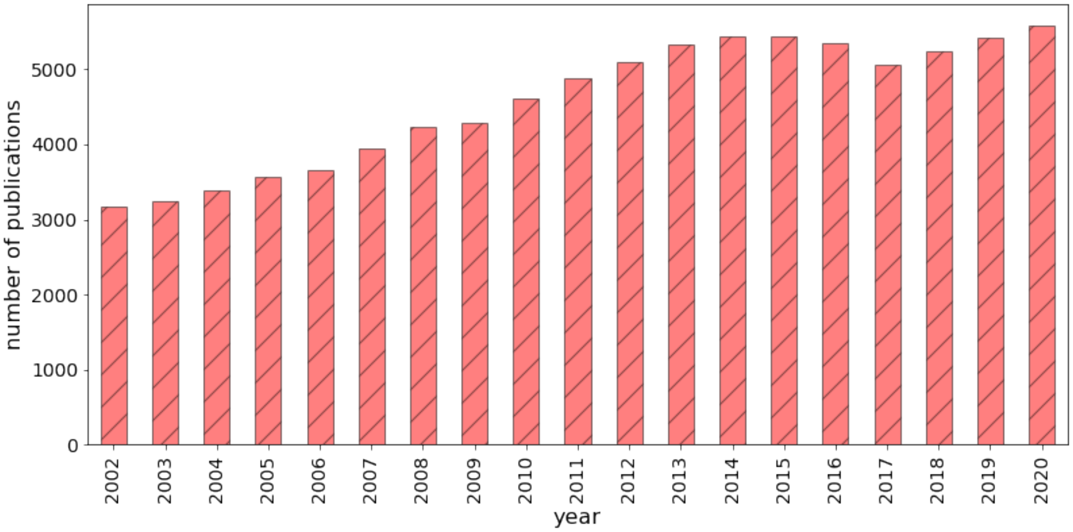


Figure 3.1. Heart disease research publications by year, 2002 - 2020.

3.2 Correlation between Heart Disease Research Funding and Publications

NHLBI grant funding for cardiovascular diseases-related publications is available via the NIH website's historical budget information disclosure [46]. Funding for heart disease research grew over 18 years period by 29% from 2,572,667 U.S. dollars in 2002 to 3,624,258 in 2020 (Figure 3.2). However, a 20.9% increase in funding occurred in the last 5 years of this time period, from 2015 to 2020. Increase in funding for 14 years prior, 2002 to and including 2015, was only 14.2%.

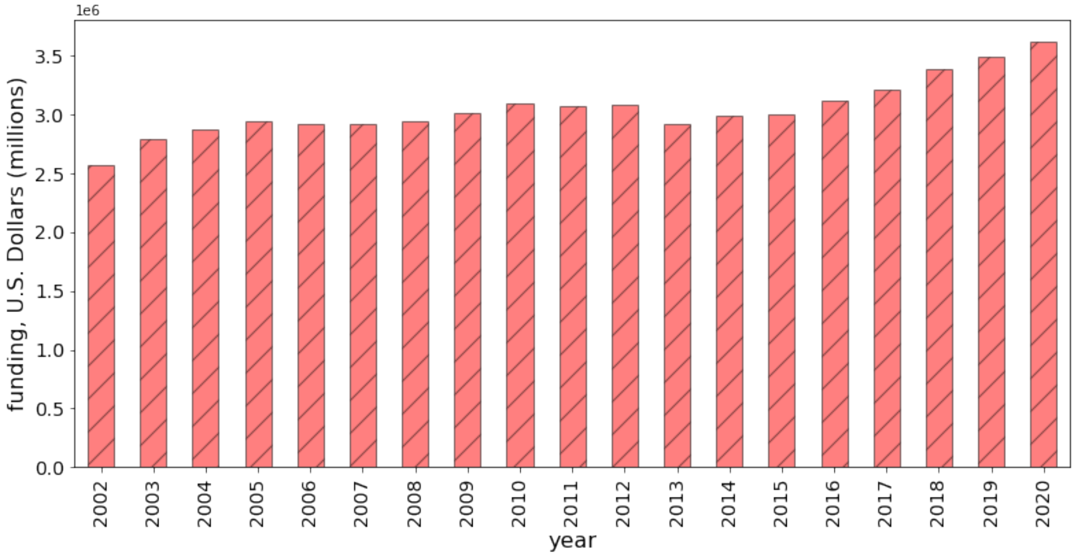


Figure 3.2. Heart Disease Research NHLBI Funding by year, 2002 - 2020.

The scatter plot between the number of publications and funding shows a moderate to good positive linear correlation with a Pearson correlation coefficient of 0.72 (Figure 3.3). The plot also shows some outliers that do not fit this relationship. These outliers fall during the time period, between 2011 and 2015, when the number of publications kept increasing despite decreasing or stagnant funding. Spending caps on scientific research funding were put in place after US Congress passed the Budget Control Act in 2011 [47]. However, there is a time lag from the time the researchers get NIH funding to the time their research gets published. This would explain the presence of these outliers in Figure 3.3. The overall trend shows that the number of heart research publications increases with increased funding.

A closer look at the distribution of NHLBI funding is shown in Figure 3.4. The distribution of data is slightly right-skewed due to the median being closer to the first

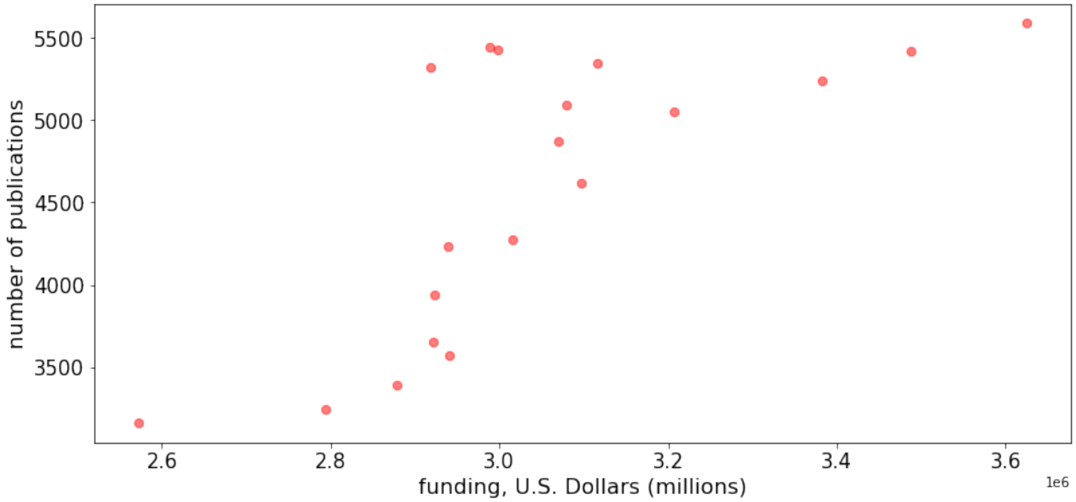


Figure 3.3. Heart disease research publications and NHLBI funding, 2002 - 2020.

quartile than to the third quartile. The data is also tightly grouped with a few upper and one lower outliers.

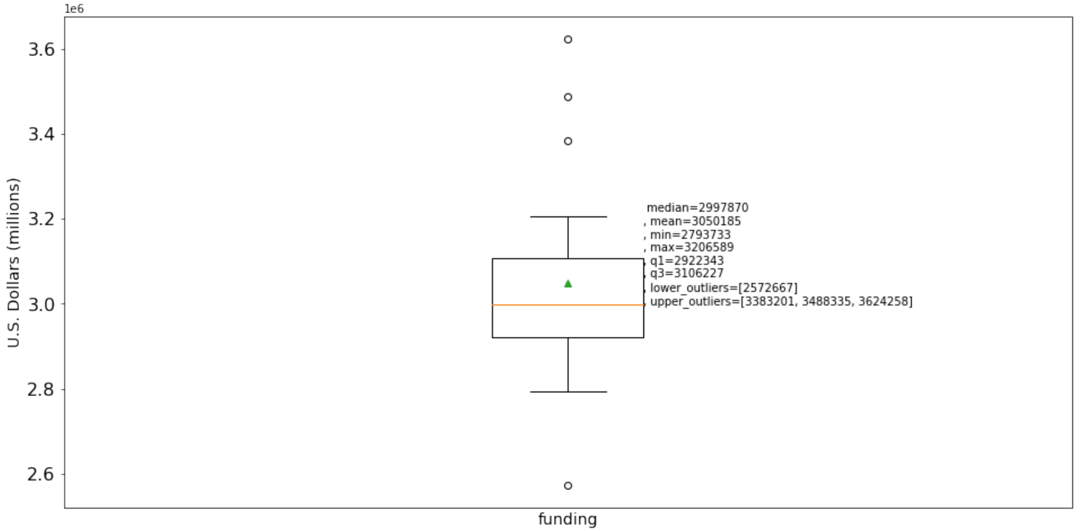


Figure 3.4. Distribution of NHLBI funding for Heart disease research publications, 2002 - 2020.

3.3 Most Frequently Publishing Journals in Heart Disease Research and Gender Disparities in Publication Opportunities

There are 2,505 individual journals in the final dataset described in Chapter 2. Out of 86,904 articles in that dataset, 21.7% (or 18,830 articles) were published in just 10 journals. These journals will be referred to as the top 10 journals for heart disease research publications going forward. Journal name for 8.0% of articles was not possible to identify in this study (Chapter 2.2). Finally, 70.4% of articles were published in the remaining journals, (Figure 3.5).

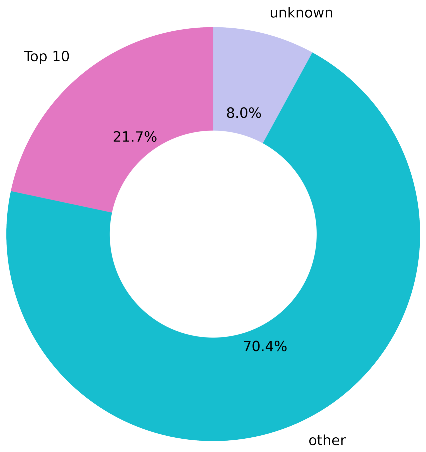


Figure 3.5. Distribution of heart disease research publications by journals, 2002 - 2020.

Out of the top 10 journals for heart disease research publications, the American Journal of Physiology, Heart and Circulatory Physiology published the most articles, 21.8% (or 4,113 articles), (Figure 3.6). Figure 3.6 also shows an overview of gender distribution for the top 10 journals. It is evident that a higher percentage of the publication's first authors are males.

These gender differences are illustrated in more detail in Figure 3.7. There are no journals in the top 10 journals for heart disease research that published more articles written by female authors than male. Among the top 10 journals, the highest percentage of articles written by female authors was published by the Journal of the American Heart Association, with 40.2% of first authors being female vs 56.4% being male. The Journal of the American College of Cardiology has the lowest number of publications with females as first authors at 26.5% vs 71.1% for males.

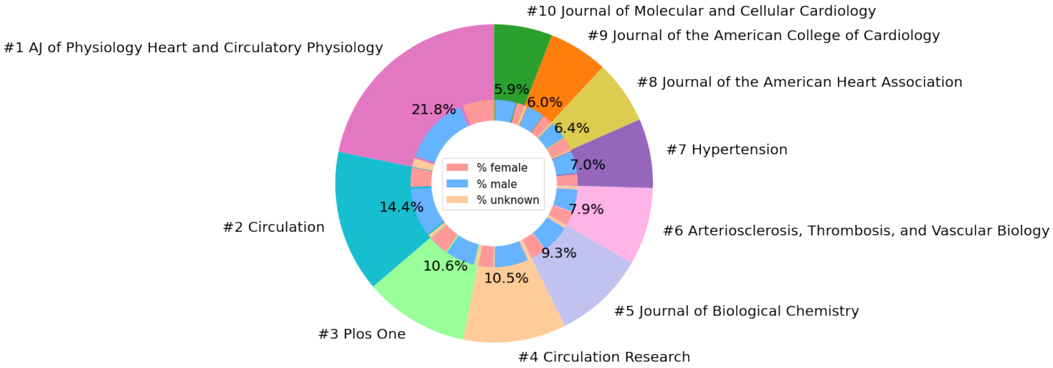


Figure 3.6. Overview of gender distribution for top 10 journals for heart disease research publications, 2002 - 2020.

Out of the top 100 journals ordered by the number of publications, only 6 journals published more female authors than male ones. The largest percentage of female authors out of these 6 journals was published by the Journal of Clinical Endocrinology and Metabolism, 56.5% female vs 37.8% male, (Figure 3.8). However, it is important to point out that in terms of raw numbers, this only represents 246 articles in total and that the Journal of Clinical Endocrinology and Metabolism is ranked 54th in the top 100 journals in this dataset.

As described in Section 3.2.3, we ranked all journals in this study in order based on the SJR score. Data shows that in the heart disease research field, journals with high ranking scores among all journals in the SCImago Journal & Country Rank database do not publish more articles than journals with lower ranking scores, (Table 3.1, 3.2). Thus, the journal with the most publications, the American Journal of Physiology-Heart and Circulatory Physiology, has a ranking order of 2,216 out of 32,942 possible. In addition, the ten journals with the highest ranking based on SJR score in this database only account for a total of 862 articles, (Table 3.2).

When the SCImago Journal & Country Rank database is filtered to only include journals in cardiology and cardiovascular medicine, the ranking order, based on the SJR score, significantly goes up. In this group of journals, the journal with the most publications, the American Journal of Physiology-Heart and Circulatory Physiology, has a ranking order of 58 out of 367 possible, (Table 3.1). Notably, 3 journals (Plos One, Journal of Biological Chemistry, and Hypertension) that are in the top 10 list in this study are not even included in this more specialized cardiology and cardiovascular medicine section of the SCImago Journal & Country Rank database, (Table 3.1).

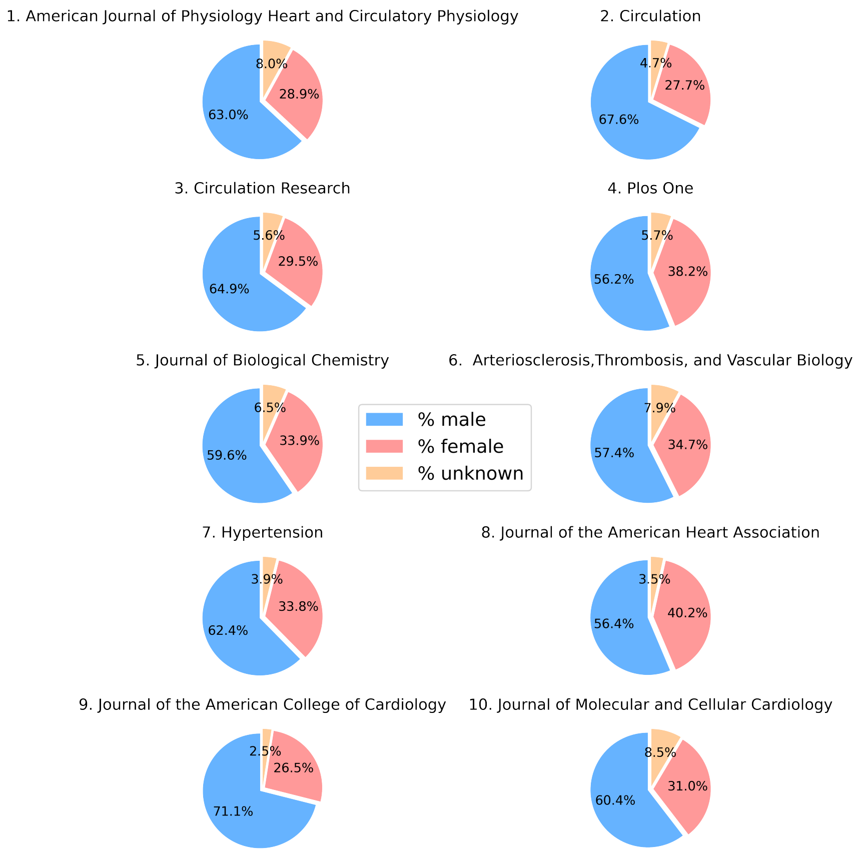


Figure 3.7. Gender distribution for top 10 journals for heart disease research publications, 2002 - 2020.

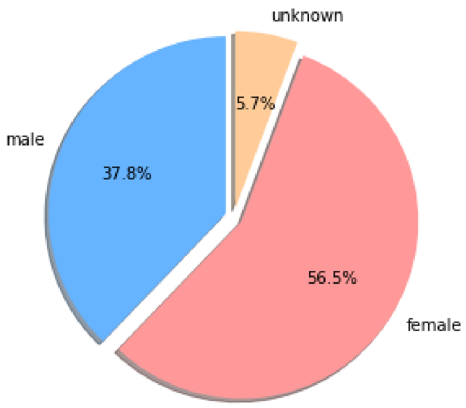


Figure 3.8. Journal, journal of clinical endocrinology and metabolism, with majority of publications with women as first authors in heart disease research, 2002 - 2020.

Table 3.1. Journal Ranking Distribution for the Top 10 Journals for Heart Disease Research Publications, 2002 - 2020.

Top 10 Order	Journal	Ranking Order based on SJR score, All Journals	Ranking Order based on SJR score, Cardio Research Journals	Number of Records
1	American Journal of Physiology-Heart and Circulatory Physiology	2,216	58	4,113
2	Circulation	134	2	2,703
3	Plos One	4,434	none	1,997
4	Circulation Research	299	8	1,979
5	Journal of Biological Chemistry	1,003	none	1,771
6	Arteriosclerosis, Thrombosis, and Vascular Biology	642	14	1,475
7	Hypertension	650	none	1326
8	Journal of the American Heart Association	916	29	1,207
9	Journal of Molecular and Cellular Cardiology	1,920	50	1,144
10	Journal of the American College of Cardiology	73	1	1,115

Table 3.2. Heart Disease Research Publications Distribution in Top Ranking Journals According to the SJR Ranking Order , 2002 - 2020.

Ranking Order based on SJR score, All Journals	Journal	Number of Records
7	Cell	116
14	New England Journal of Medicine	182
17	Nature Medicine	161
18	Nature Methods	19
19	Nature Genetics	150
27	Nature	174
32	Nature Biotechnology	22
35	Nature Materials	11
37	Nature Nanotechnology	4
38	Immunity	23

3.4 Most Frequently Publishing Research Institutions in Heart Disease Research and Gender Disparities in Publication Opportunities

There are 4,053 individual research institutions in the final dataset described in Chapter 2. Out of 86,904 articles in that dataset, 16.9% (or 14,675 articles) were published by first authors affiliated with the top 10, based on the number of publications, and institutions. We will refer to these institutions as the top 10 institutions for heart disease research publications going forward. The institution name for 0.1% of institutions was not possible to identify in this study (Section 2.2). Finally, 83.0% of articles were produced in the remaining institutions, (Figure 3.9).

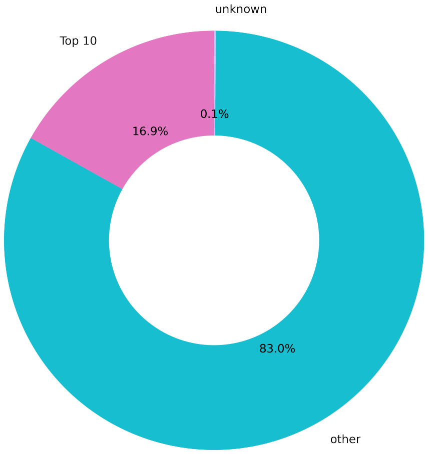


Figure 3.9. Distribution of heart disease research publications by research institutions, 2002 - 2020.

Out of the top 10 institutions for heart disease research publications, John Hopkins University-affiliated researchers published the most articles, 11.7% (or 1,724 articles), (Figure 4.2). Figure 4.2 also shows an overview of the gender distribution for the top 10 institutions. It is evident that a higher percentage of the publications' first authors are males.

These gender differences are illustrated in more detail in Figure 3.11. There are no institutions in the top 10 institutions for heart disease research that published more articles written by female authors than male authors. Among the top 10 institutions, the highest percentage of articles written by female authors was published by Columbia

University, with 38.6% of first authors being female vs 55.3% being male. Massachusetts General Hospital has the lowest number of publications with females as first authors at 24.2% vs 70.8% for males.

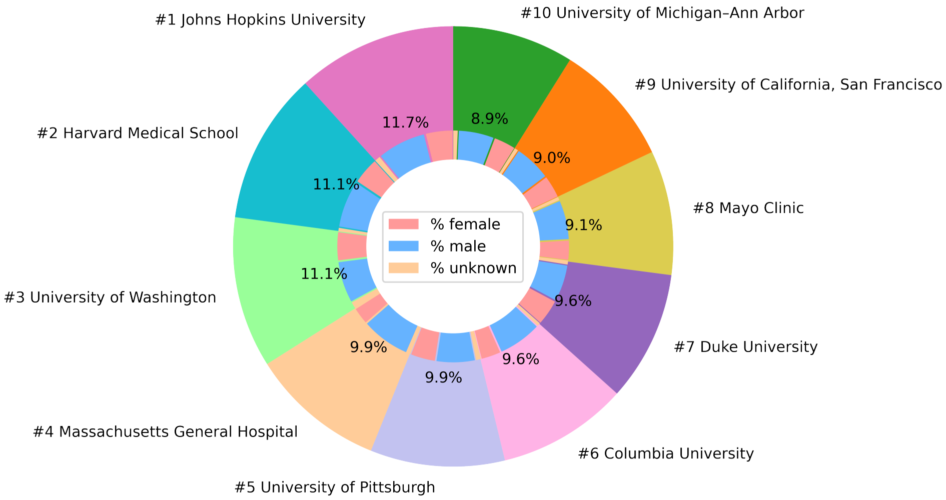


Figure 3.10. Overview of gender distribution for top 10 institutions for heart disease research publications, 2002 - 2020.

Out of the top 100 institutions ordered by the number of publications, no institutions had more female-authored articles than male ones. Only at rank 104, Tufts University, there are more female first authors than male, 57% female vs 37% male, (Figure 3.12). However, it is important to point out that in terms of raw numbers Tufts University only published 156 publications in total. Overall, 1,262 out of 4,053 research institutions had a majority, greater than 50%, of publications first-authored by female authors.

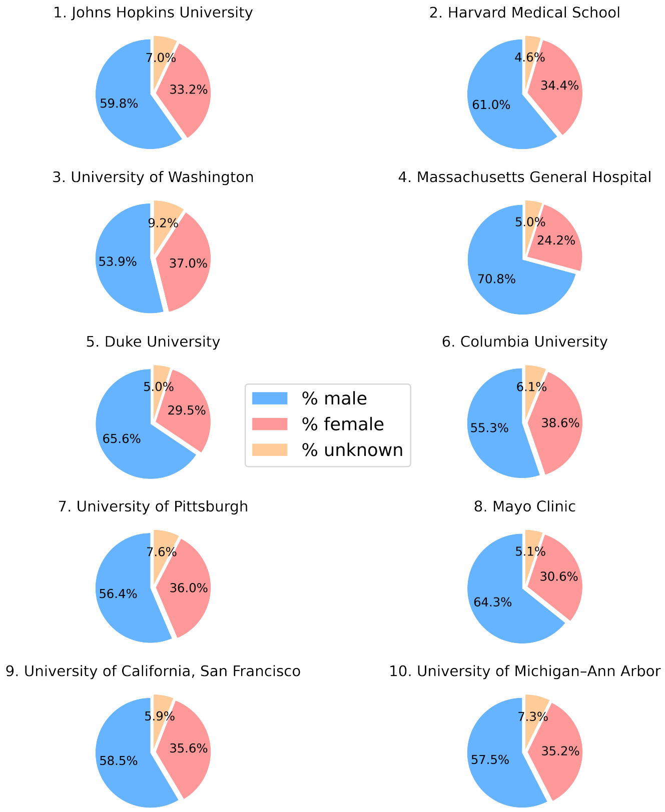


Figure 3.11. Gender distribution for top 10 institutions for heart disease research publications, 2002 - 2020.

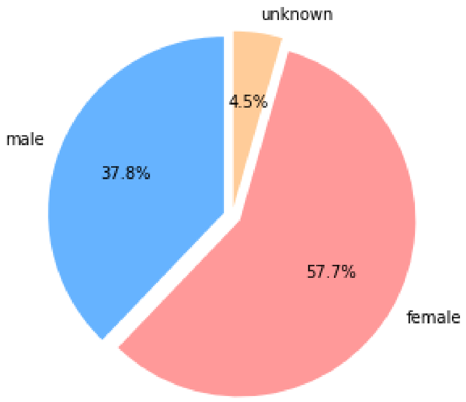


Figure 3.12. Institution, Tufts University, with a majority of publications with women as first authors in heart disease research, 2002 - 2020.

3.5 Gender Disparity in Citation Practices in Heart Disease Research

The number of total citations for heart disease research publications remained stable over a 10 year period, from 250,566 in 2002 to 249,134 in 2011 (Figure 3.13). However, number of total citations slowly decreased to 61,953 by 2020. This is expected, as these are more recent publications.

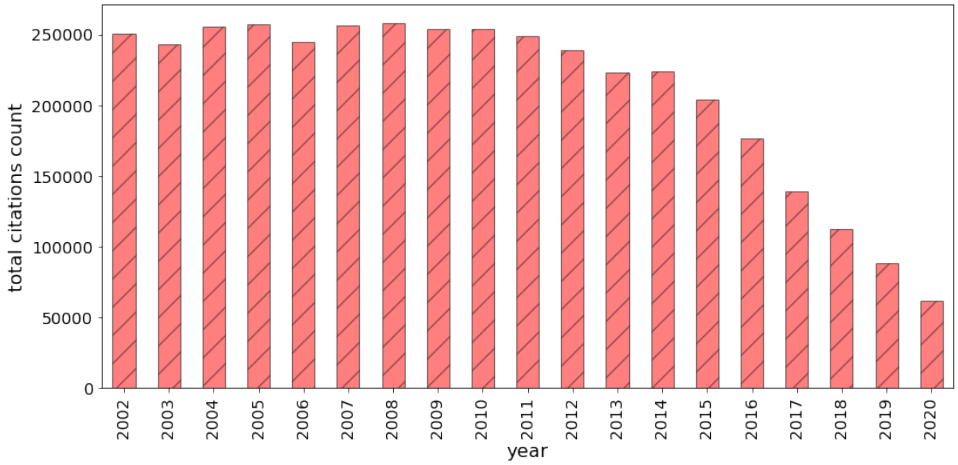


Figure 3.13. Total citations for heart disease research publications, 2002 - 2020.

Overall, publications where the first author was male were cited twice as much compared to publications where a woman was the first author, (Figure 3.14).

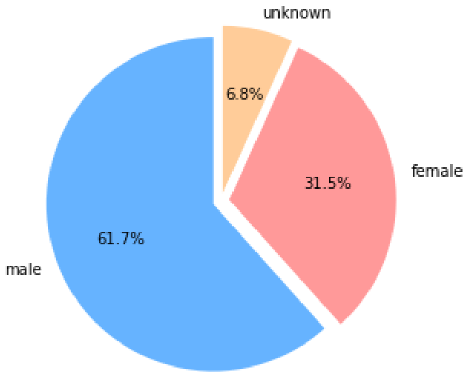


Figure 3.14. Total citations for heart disease research publications by first author's gender, 2002 - 2020.

Gender differences as it relates to the citations, represented by NIH Percentile, are further illustrated in Figure 3.15. We included only papers with NIH Percentile greater or equal to 5% in these box plots. This was done to avoid skewed data due to the large number of more recent papers that do not have a lot of citations yet. The distribution of data is symmetrical and tightly grouped for both genders with some upper and lower outliers. The box plot for female authors shows slightly more compact data than for male authors. The mean for NIH Percentile for papers written by male authors is 1.8 times higher than for papers with female first authors. For males, the interquartile range and the lengths of the whiskers are slightly larger than for females. This suggests that NIH Percentile is dispersed over a larger range for male authors compared to female authors.

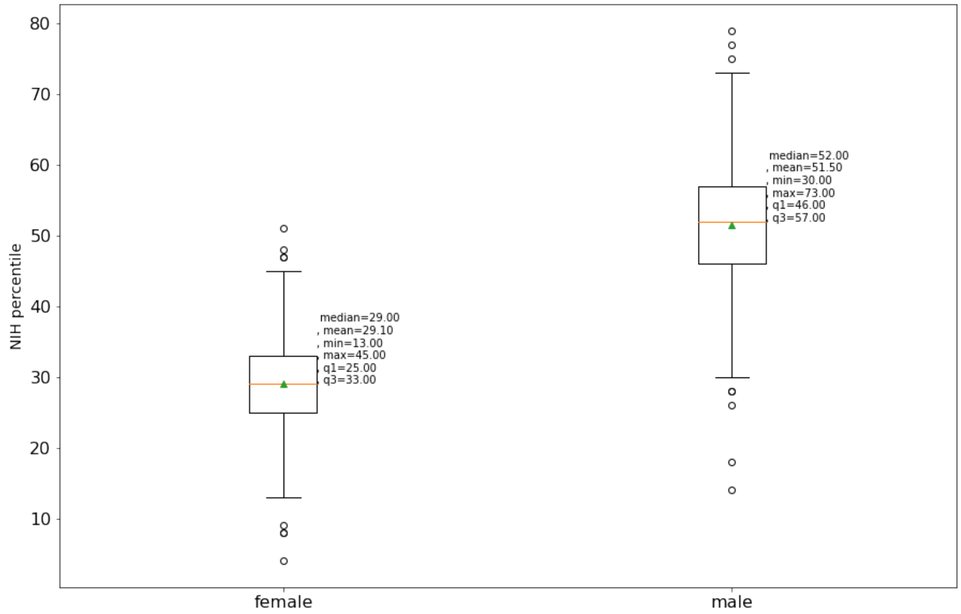


Figure 3.15. Distribution of citations, as measured by NIH percentile, for heart disease research publications by first author's gender, 2002 - 2020.

The top 10 journals, when ranked by the number of total citations, account for 26% or 652,120 citations out of 3,993,893 total. Figure 3.16 shows an overview of gender distribution for these top 10 journals. It is evident that articles written by men have a higher percentage of total citations.

When journals are ordered based on the article's first author's gender, the order of the journals looks slightly different for female, male, or both authors. Table 3.3 illustrates this for the top 10 journals in each category.

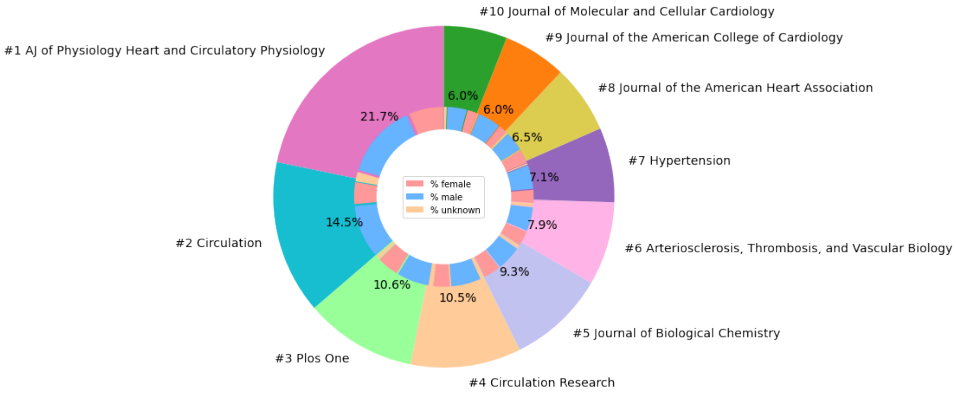


Figure 3.16. Overview of gender distribution for top 10 journals with most citations for heart disease research, 2002 - 2020.

The top 10 research institutions, when ranked by the total number of citations, account for 12.8% or 514,704 citations out of 3,993,893 total. Figure 3.17 shows an overview of gender distribution for these top 10 institutions. It is evident that articles written by men have a higher percentage of the total citations.

When research institutions are ordered based on the article's first author's gender, the order of the institutions looks slightly different for female, male, or both authors. Table 3.4 illustrates this for the top 10 journals in each category.

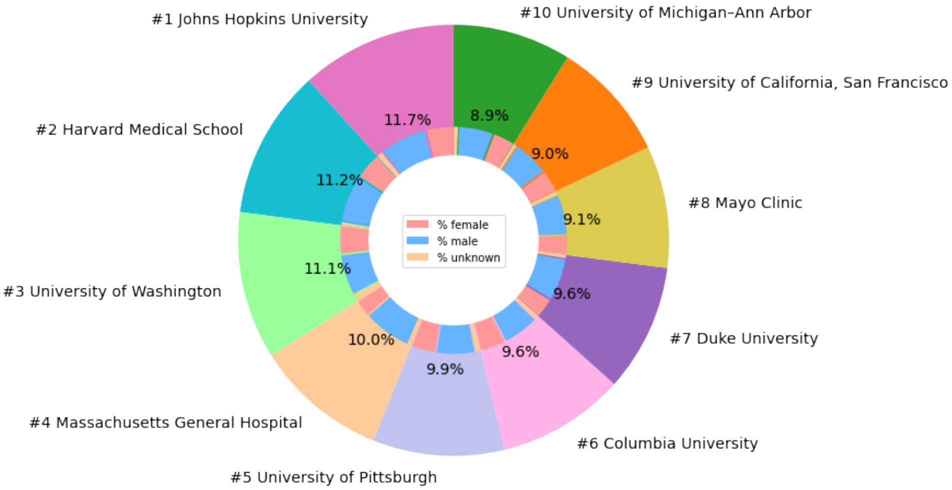


Figure 3.17. Overview of gender distribution for top 10 institutions with most citations for heart disease research, 2002 - 2020.

Table 3.3. Gender Differences in Top 10 Journals with Most Citations for Heart Disease Research Publications, 2002 - 2020.

Top 10 Order	Both Genders	Female	Male
1	American Journal of Physiology-Heart and Circulatory Physiology, 21.7%	American Journal of Physiology-Heart and Circulatory Physiology, 19.8%	American Journal of Physiology-Heart and Circulatory Physiology, 21.9%
2	Circulation, 14.5%	Plos One, 12.7%	Circulation, 15.7%
3	Plos One, 10.6%	Circulation, 12.7%	Circulation Research, 10.9%
4	Circulation Research, 10.5%	Journal of Biological Chemistry, 9.9%	Plos One, 9.5%
5	Journal of Biological Chemistry, 9.3%	Circulation Research, 9.7%	Journal of Biological Chemistry, 8.8%
6	Arteriosclerosis, Thrombosis, and Vascular Biology, 7.9%	Arteriosclerosis, Thrombosis, and Vascular Biology, 8.6%	Arteriosclerosis, Thrombosis, and Vascular Biology, 7.2%
7	Hypertension, 7.1%	Journal of the American Heart Association, 8.2%	Hypertension, 7.0%
8	Journal of the American Heart Association, 6.5%	Hypertension, 7.5%	Journal of the American College of Cardiology, 6.8%
9	Journal of the American College of Cardiology, 6.0%	Journal of Molecular and Cellular Cardiology, 5.9%	American Journal of Cardiology, 6.4%
10	Journal of Molecular and Cellular Cardiology, 6.0%	Journal of the American College of Cardiology, 5.0%	Journal of American Heart Association, 5.8%

Table 3.4. Gender Differences in Top 10 Institutions with Most Citations for Heart Disease Research Publications, 2002 - 2020.

Top 10 Order	Both Genders	Female	Male
1	John Hopkins University, 11.7%	University of Washington, 12.0%	Massachusetts General Hospital, 11.7%
2	Harvard Medical School, 11.2%	John Hopkins University, 11.5%	John Hopkins University, 11.6%
3	University of Washington, 11.1%	Harvard Medical School, 11.3%	Harvard Medical School, 11.3%
4	Massachusetts General Hospital, 10.0%	Columbia University, 10.9%	Duke University, 10.4%
5	University of Pittsburgh, 9.9%	University of Pittsburgh, 10.5%	University of Washington, 9.9%
6	Columbia University, 9.6%	University of California, San Francisco, 9.5%	Mayo Clinic, 9.7%
7	Duke University, 9.6%	University of Michigan-Ann Arbor, 9.2%	University of Pittsburgh, 9.2%
8	Mayo Clinic, 9.1%	University of North Carolina at Chapel Hill, 8.7%	Stanford University, 8.8%
9	University of California, San Francisco, 9.0%	Duke University, 8.3%	Columbia University, 8.8%
10	University of Michigan-Ann Arbor, 8.9%	Mayo Clinic, 8.2%	University of California, San Francisco, 8.7%

3.6 Gender Gap in Publication Opportunities in Heart Disease Research

Overall, over the time period of 18 years, from 2002 - 2020, there were 1.7 times more publications where the first author was male compared to publications where a woman was the first author, (Figure 3.18).

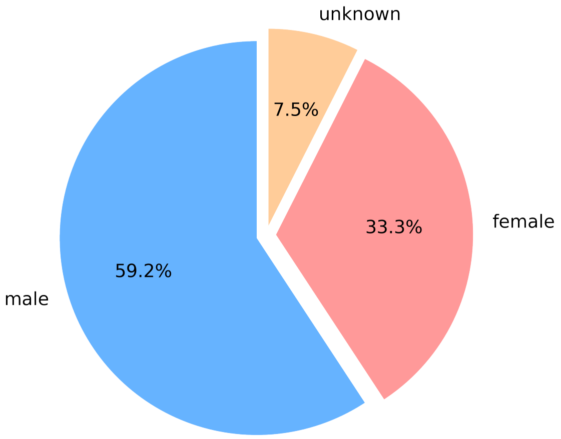


Figure 3.18. Heart disease research publications by first author's gender, 2002 - 2020.

The number of publications for both genders grew steadily from 2002 to 2020, (Figure 3.19). However, at no time there were more publications written by women than men. The gender gap in the number of publications remained almost the same, with 1,180 in 2002 and 1,041 in 2020.

The ratio of male to female first authors, normalized by the total (male, female, and unknown combined) number of publications, slightly decreased from 2.4 in 2002 to 1.5 in 2020, (Figure 3.20).

The distribution of heart disease research publications by first author's gender is further illustrated in Figure 3.21. The distribution of data is asymmetrical and tightly grouped for both genders. The mean for the number of papers written by male authors is 1.8 times higher than for papers with female first authors. For males, the interquartile range and the lengths of the whiskers are slightly larger than for females. This suggests that the number of publications is dispersed over a larger range for female authors compared to male.

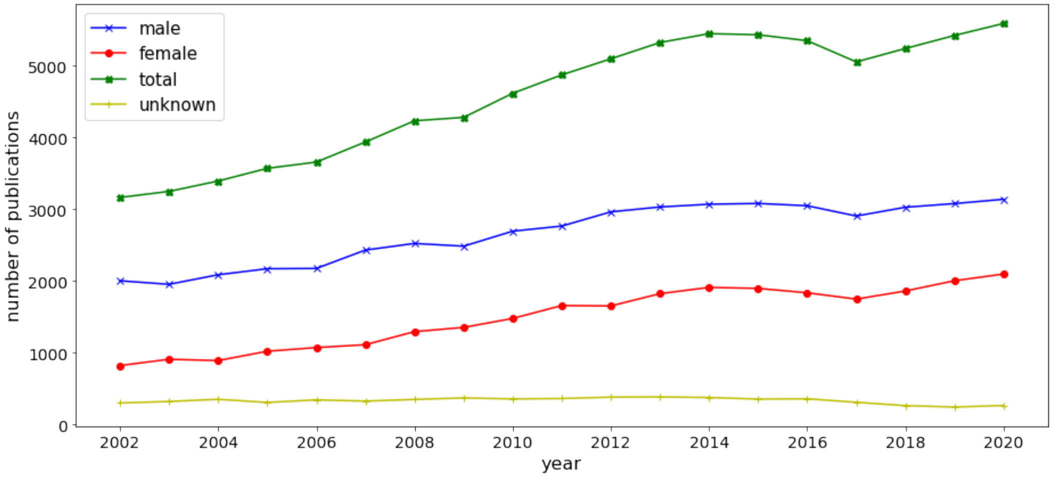


Figure 3.19. Heart disease publications by first author's gender and year, 2002 - 2020.

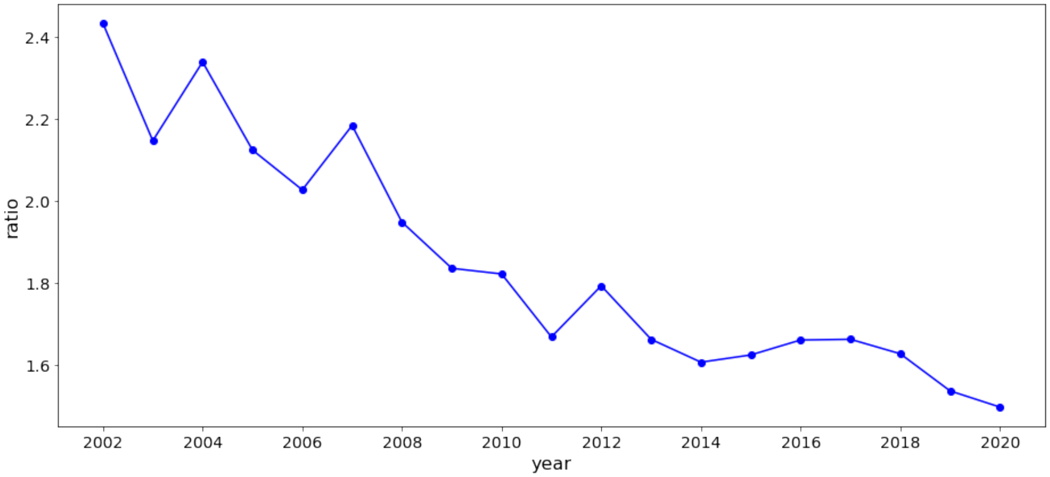


Figure 3.20. Change in ratio of male to female first authors of heart disease research publications, 2002 - 2020.

The gender gap is also apparent when the number of publications is plotted against the journal ranking order based on SJR score (Figure 3.22). Due to the very large number of journals in the study, we grouped journals into bins of 100 for this visualization. The gender gap is larger in higher-ranking journals.

An overview of gender differences in who publishes the majority of papers among the first authors of heart disease research publications is presented in Table 3.5. Out of 2,505 journals registered in the database in this study, half as many journals published more articles written by men (first author) compared to women as first authors (1,558 vs 832). The gender difference is even more pronounced in the top 100

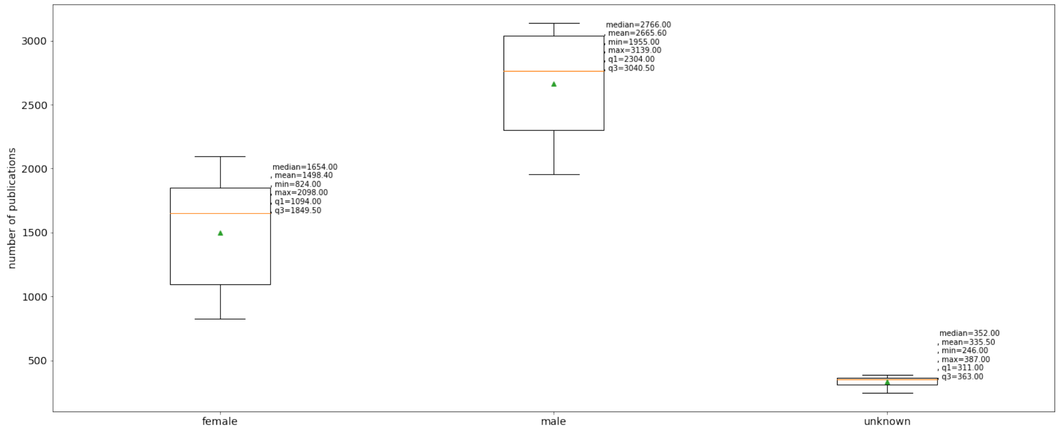


Figure 3.21. Distribution of heart disease research publications by first author's gender, 2002 - 2020.

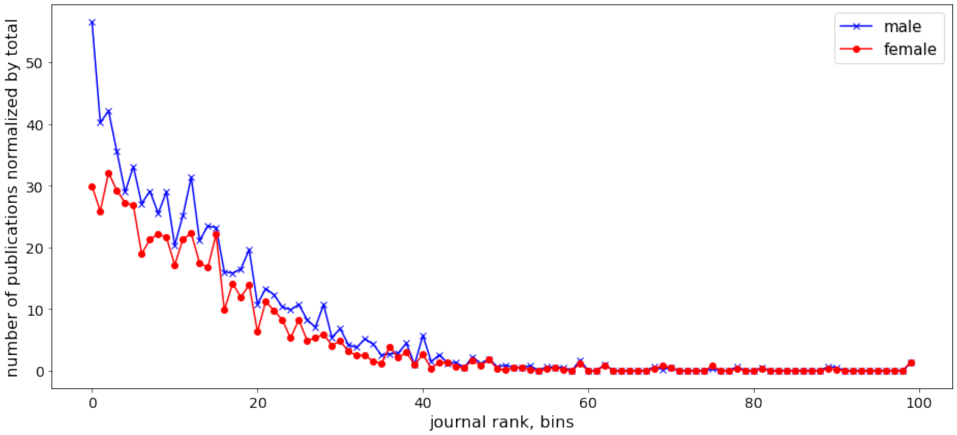


Figure 3.22. Gender gap in the number of heart disease research publications by journal ranking, 2002 - 2020.

journals for heart disease research as established by this study. Thus, only 6 journals in the top 100 had published more articles with women as first authors than men.

Similar trends are seen for the research institutions' article's first author associated with, Table 3.5. Out of 4,053 research institutions registered in the database in this study, half as many journals published more articles written by men (first author) compared to women as first authors (2,756 vs 1,262). Noteworthy is that in the top 100 institutions for heart disease research as established by this study, there was not a single research institute where female researchers published more papers, as first authors, than men.

Table 3.5. Overview of Gender Differences: Majority of First Authors of Heart Disease Research Publications by Journals and Institutions, 2002 - 2020.

Feature	Female	Male	Unknown	Total
Number of all journals with 1st author's publications >= 50%	832	1558	206	2505
Number of top 100 journals with 1st author's publications >= 50%	6	89	1	100
Number of all institutions with 1st author's publications >= 50%	1262	2756	252	4053
Number of top 100 institutions with 1st author's publications >= 50%	0	93	0	100

CHAPTER 4
FACTORS ASSOCIATED WITH IMPACTFUL SCIENTIFIC PUBLICATIONS

The purpose of this chapter is to identify key factors that contribute to successful research outcomes in estimating the impact of cardiovascular research. One important indicator of a research paper's impact and influence in the field is the number of citations it receives. We use NIH Percentile as an estimate of the number of people benefiting from a publication. This chapter provides an overview of the target and feature selection, data pre-processing steps, the machine learning models employed, and an evaluation of the results obtained.

4.1 Target and Features Selection

We used Machine Learning modeling in this study to answer the question of if it is possible to predict the paper's success based only on the information available at the time of submission. We choose NIH Percentile, a rank of any individual paper's RCR score relative to all other NIH publications, as a measure of publication success. Therefore, NIH Percentile is considered a dependent target variable in ML modeling.

Features are independent variables in a ML model. Several points are important to consider when performing feature selection. First, features that could be considered proxies or duplicate targets need to be avoided to prevent feature leakage. With this in mind, we did not use the total citation count, citations per year, field citation ratio, RCR score, and some other citation-related data as features for ML modeling in this study.

Second, considering the predictive goal of the ML model, we only considered data that would be available at the moment of paper submission to a journal in the feature selection.

Third, to determine the strength of the linear relationship between features (independent variables) and target (dependent variable) and also to identify the multi-collinearity of features, we built the correlation matrix for raw data by using the function *heatmap*() from the seaborn library (Figure 4.1) [48], [49]. The matrix shows Pearson correlation coefficient values. There was no strong linear association between the target, NIH Percentile, and any of the features. Moderate correlation exists between the target and the Journal feature and between the Journal and Journal Rank features. The rest of the correlations were weak or no linear correlation was found at all. We selected a total of 11 features to be used in ML modeling (Table 4.1).

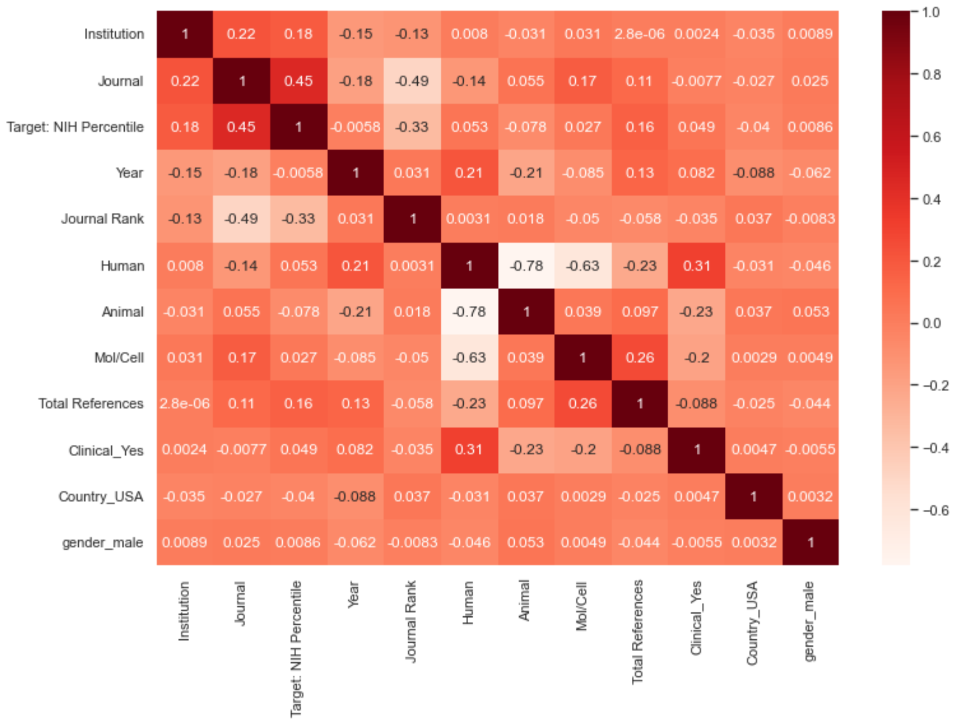


Figure 4.1. Correlation heatmap for raw data.

4.2 Data Pre-processing

Before performing ML modeling both target and features, were pre-processed as follows:

4.2.1 Categorical Features

For the Country feature, we split all countries into two categories, USA and not USA. Since the majority of records (64,616) had the first author's country as USA vs. not USA (8,095), we combined all other countries into one group to create two binary categories. However, it is important to notice that these two categories are imbalanced.

We encode binary features, Gender, and Country, via one-hot encoding. For the feature Gender, two categories were male and female.

We used total citation counts for each record to target-encode non-numeric features, namely Institution and Journal. We analyzed the total citations data for

Table 4.1. Features Selected for Machine Learning Regression Models.

Feature	Description
Year	The year an article was published
Institution	Research institution first author affiliated with
Journal	Name of the journal where the article was published
Journal Rank	Ranking number of the journal as determined by SCImago Journal and Country Ranking
Human biology scores	Fraction of Medical Subject Headings (MeSH) terms that are in the human category out of all MeSH terms assigned to the article. MeSH terms are assigned to PubMed articles to provide information about the content of the publications. MeSH terms that can be used during PubMed database search.
Animal biology scores	Fraction of MeSH terms that are in the animal category out of all MeSH terms assigned to the article.
Mol/Cell biology scores	Fraction of MeSH terms that are in the molecular/cellular category out of all
Clinical	Paper meets the definition of a clinical article
Total references	Number of references on the article's reference list
Country	Country of a research institution first author affiliated with
Gender	First author's gender

skewness with the help of the *skew*() function from the Pandas library and transformed it using the log function to reduce skewness before target-encoding [42].

4.2.2 Numeric Features

We analyzed each numeric feature for skewness. A distribution is considered highly skewed for values greater than 1 or less than -1 and moderately skewed for values between 0.5 and 1 or -0.5 and -1. Values between -0.5 and 0.5 indicate that the distribution is fairly symmetrical [50].

For each quantitative feature, we perform feature transformations to decrease skewness including log, sin, power of 2, and power of 3 transformations. After the transformations, we analyzed features for skewness again and, for each feature, chose the least skewed from the transformations of the feature and the original feature.

4.2.3 General Pre-Processing

We removed all records with unknown features resulting in 72,711 records (a 16% reduction from the original total of 86,904 records). We described the reasons for records with unknown information in Section 2.2 and Table 2.3. The features with the unknown entries can introduce bias in the results, decrease the accuracy and statistical power of the model, and would be misleading in the ML model training [51].

To re-evaluate the strength of the relationship between features and target and also to identify the multi-collinearity of features, we built the correlation matrix for the pre-processed data (Figure 4.2). There is no strong linear relationship between the target, NIH Percentile, and any of the features. After data pre-processing, the correlation between the features Journal and Journal Rank became stronger (Figure 4.2) compared to the one for the raw data (Figure 4.1). The rest of the correlations between features remained either weak or no linear correlation was found.

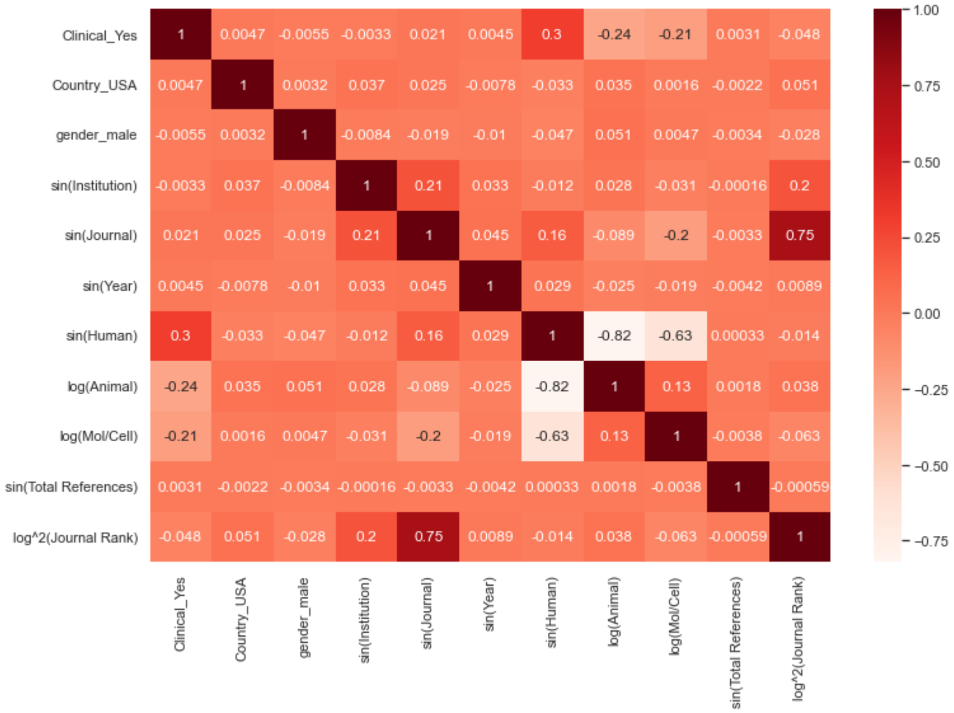


Figure 4.2. Correlation heatmap for pre-processed data.

4.2.4 Train-Test Split

We split the dataset of 72,711 records at an 80:20 ratio (58,168:14,543 records) into train and test sets. We normalized the train and test data sets with

$MinMaxScaler()$ from Scikit-learn [52]. $MinMaxScaler()$ transforms features by scaling each feature to a 0 to 1 range. This includes features and the target.

4.3 Machine Learning Modeling

The objective of this section is to use supervised modeling techniques to predict a research publication's impact in the cardiovascular research field. Regression and classification algorithms are two models that we used to predict the success of the publication.

4.3.1 Regression

Initially, we used the following state-of-the-art supervised regression modeling techniques to predict the continuous target values.

Random Forest Regression (RFR) outputs an average value from simultaneously made predictions from multiple random decision trees. The model uses a different subset of features for each decision tree, which is also chosen at random. The depth of these trees is shallow because only a subset of features is used. Some of the pros of RFR are reduced overfitting and suitability for both categorical and continuous values [53]. Random Forest is also not sensitive to outliers [54]. However, the model is complex and requires a lot of computational power to train the data and tune the hyperparameters [53].

Gradient Boosting Regression (GBR) outputs a prediction made from sequentially constructed multiple decision trees with each new tree improving on the prediction of the previous one. Unlike Random Forests, Gradient Boosting is sensitive to outliers and to overfitting when too many trees are used [55]. However, Gradient Boosting is a slow learning model, which is a more robust and generalized model leading to more accurate predictions. Also, like Random Forests, it is suitable for both categorical and continuous values [55].

Extreme Gradient Boosting Regression (XGBR) is a more regularized form of GBR, which helps prevent a model from overfitting by controlling model complexity. To control overfitting, the XGBR model uses L1 and L2 (Lasso and Ridge Regressions correspondingly) regularization techniques, which apply extra penalty terms to the loss function [56]. Additionally, XGBR is faster than GBR because the process can be ”parallelized across clusters” [57].

Support Vector Machine (SVM) finds the best-fit line, a hyperplane, that fits a maximum number of data points. A SVM model does not need all of the data points for training. Moreover, it prevents overfitting by using a regularization term, which helps "balance complexity and accuracy of the model" [58]. However, SVM is computationally demanding for a large data set [58].

Multi-layer Perceptron (MLP) is an artificial feedforward neural network that has at least 3 layers (input, hidden, output) of neural network units that do computations on the input data [59]. MLP uses backpropagation during model training to reduce the error and increase the accuracy of prediction [60]. A MLP model is fast, can process a large amount of data, and can solve complex nonlinear problems [60]. However, the model is computationally expensive and requires sufficient training to perform accurately [60].

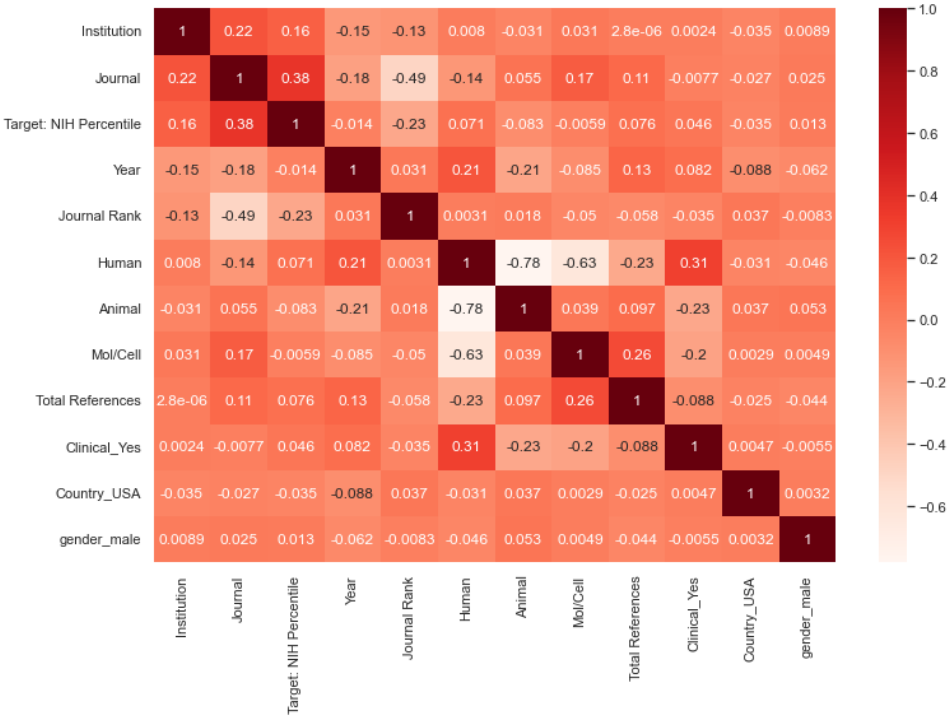
Model Stacking (MS) combines outputs from different ML models, GBR, and RFR models in this study, and then runs it via a so-called meta-learner model, which in this study is a Ridge Regression (RR) model [61]. RR is a regression model that is used for the regularization of data that has multi-collinearity among independent variables. The MS technique helps to create robust models that work well on unseen test data [61]. However, one of the disadvantages of MS is increased computational time, as each of the models has to be trained on the training dataset.

We used the RFR, GBR, SVM, and MLP model packages from Scikit-learn library [52]. The XGBR model is from the XGBoost library [62]. *RandomizedSearchCV*() from Scikit-learn with 5-fold cross-validation was used to tune the hyperparameters of the RFR, GBR, and XGBR models [52].

4.3.2 Classification

Classification is a supervised learning algorithm employed for predicting discrete class labels. For classification modeling the target variable, NIH Percentile, was converted from a continuous variable into a categorical variable with two discrete classes: publications with NIH Percentile equal to or above 80% and publications with NIH Percentile below 80%. NIH-funded papers with RCR 1.0 perform better in terms of citation statistics than 50% of NIH-funded papers [24]. 80% NIH Percentile corresponds to an RCR score of 2.39 or in other words, the article is about 2.39 times as influential as an average article with NIH Percentile of 50%. For reference, 99.9% NIH Percentile represents an RCR score of 38.0. We choose 80% NIH Percentile as a cut-off point for two classes after testing other cut-off points (50, 60, 70, and 90 NIH Percentile) with all of the classification models described in section 4.3. 50, 60, and 70 NIH Percentile cut-off points all resulted in poorer predictions than 80%. 90% NIH Percentile cut-off resulted in better prediction accuracy than 80%, but it represents only 10% of all records compared to 21% for 80% NIH Percentile and therefore we decided that it was too selective and exclusive for a fair prediction of a publication's success.

To determine the strength of the linear relationship between features and the target variable and also to identify the multi-collinearity of features, we generated a correlation matrix for raw data for the classification modeling approach (Figure 4.3) [48]. No strong linear association between the target, NIH Percentile, and any of the features was found. The rest of the data pre-processing was done as described (Section 4.2) with one additional step. Synthetic Minority Oversampling Technique (SMOTE) from the imbalanced-learn library was applied to the data to balance the minority class and to control oversampling [63]. This resulted in an additional 42,325 records generated by SMOTE and a final dataset of 115,036 records. The dataset was subsequently split at an 80:20 ratio (92,028:23,008 records) into train and test sets.



	Institution	Journal	Target: NIH Percentile	Year	Journal Rank	Human	Animal	Mol/Cell	Total References	Clinical_Yes	Country_USA	gender_male
Institution	1	0.22	0.16	-0.15	-0.13	0.008	-0.031	0.031	2.8e-06	0.0024	-0.035	0.0089
Journal	0.22	1	0.38	-0.18	-0.49	-0.14	0.055	0.17	0.11	-0.0077	-0.027	0.025
Target: NIH Percentile	0.16	0.38	1	-0.014	-0.23	0.071	-0.083	-0.0059	0.076	0.046	-0.035	0.013
Year	-0.15	-0.18	-0.014	1	0.031	0.21	-0.21	-0.085	0.13	0.082	-0.088	-0.062
Journal Rank	-0.13	-0.49	-0.23	0.031	1	0.0031	0.018	-0.05	-0.058	-0.035	0.037	-0.0083
Human	0.008	-0.14	0.071	0.21	0.0031	1	-0.78	-0.63	-0.23	0.31	-0.031	-0.046
Animal	-0.031	0.055	-0.083	-0.21	0.018	-0.78	1	0.039	0.097	-0.23	0.037	0.053
Mol/Cell	0.031	0.17	-0.0059	-0.085	-0.05	-0.63	0.039	1	0.26	-0.2	0.0029	0.0049
Total References	2.8e-06	0.11	0.076	0.13	-0.058	-0.23	0.097	0.26	1	-0.088	-0.025	-0.044
Clinical_Yes	0.0024	-0.0077	0.046	0.082	-0.035	0.31	-0.23	-0.2	-0.088	1	0.0047	-0.0055
Country_USA	-0.035	-0.027	-0.035	-0.088	0.037	-0.031	0.037	0.0029	-0.025	0.0047	1	0.0032
gender_male	0.0089	0.025	0.013	-0.062	-0.0083	-0.046	0.053	0.0049	-0.044	-0.0055	0.0032	1

Figure 4.3. Correlation heatmap: features and target for raw data.

We used the following state-of-the-art supervised classification models to predict the NIH Percentile categories.

Random Forest Classifier (RFC) has the same algorithm as the Random Forest model in 4.3.1, but instead of average prediction value, RFC outputs the majority class from simultaneously made predictions from multiple random decision trees [64].

Gradient Boosting Classifier (GBC) is similar to its regression counterpart. It works by creating multiple decision trees in sequence, with each new tree improving on the prediction of the previous one. However, unlike RFR which uses Mean Squared Error as its loss function, log-likelihood is used as a loss function in GBC [65].

Extreme Gradient Boosting Classifier (XGBC) has the same qualities as described for the Extreme Gradient Boosting algorithm for regression in Section 4.3.1 with a difference that it predicts categorical values.

Logistic Regression Classifier (LRC) uses a sigmoid function to predict output for a target class [66]. LRC is an efficient algorithm that does not need a lot of computational resources. However, the simplicity of the model allows it to be outperformed by more complex models such as Random Forest and Gradient Boosting Classifiers [66].

K-Neighbors Classifier (KNN) uses the proximity of data points to predict output for a target class [67]. Since only distances between data points are being measured in the KNN model, it is very simple to execute and new data can be added at any time [68]. However, computational cost increases for larger data sets. Moreover, the KNN model does not perform accurately for datasets with a large number of features [68].

We used the RFC, GBC, LRC, and KNN model packages from the Scikit-learn library [52]. The XGBC model is from the XGBoost library [62]. $RandomizedSearchCV()$ from Scikit-learn with 5-fold cross-validation was used to find the best set of hyperparameters for all models [52].

4.4 Evaluations

We outlined three goals to evaluate the ML models' performance:

1. Evaluating the effectiveness of regression models in predicting NIH Percentile.
2. Evaluating the effectiveness of classification models in predicting success category.
3. Identifying important features.

4.4.1 Evaluating the effectiveness of regression models in predicting NIH Percentile

We evaluated the effectiveness of the regression models in predicting the NIH Percentile using the following metrics:

R squared (R^2) is the coefficient of determination that measures how much the variation in the dependent variable (target) can be attributed to variations of independent variables (features). R^2 ranges from 0 to 1, with higher numbers indicating the better fit. A value of 0 means that the model's fit is no better than random, while a value of 1 indicates a perfect fit.

Accuracy is a percent representation of the R^2.
Mean Absolute Error (MAE) is the average absolute value of the difference between the predicted value and the actual value of the dependent variable target. The lower the MAE, the better the model's accuracy.
Mean Squared Error (MSE) is the average of squared differences between the predicted value and the actual value of the dependent variable. The lower the MSE, the better the model's accuracy.
Root Mean Squared Error (RMSE) is the square root of the average of squared differences between the predicted value and the actual value of the dependent variable. The lower the RMSE, the better the model's accuracy.
Max Error (ME) is the worst-case scenario or the absolute value of the most significant difference between the predicted and actual value of a dependent variable.

The results of an evaluation of ML models are summarised in Table 4.2. None of the models perform better than a random model. Accuracy scores calculated using R^2 values were low for all the models even for the training data set. RFR, XGBR, and MS models also showed overfitting as the accuracy for the training set was approximately doubled compared to the accuracy for the testing set.

Table 4.2. Comparing Machine Learning Regression Models.

Model	Set	Accuracy %	R^2	MAE	MSE	RMSE	ME
RFR	train	52.2	0.522	0.157	0.036	0.189	0.820
	test	27.9	0.279	0.193	0.054	0.232	0.809
GBR	train	31.0	0.310	0.189	0.052	0.228	0.836
	test	28.3	0.283	0.192	0.054	0.232	0.764
XGBR	train	68.4	0.684	0.123	0.024	0.154	0.684
	test	28.4	0.284	0.190	0.054	0.232	0.761
SVM	train	28.1	0.281	0.194	0.054	0.232	0.780
	test	24.3	0.243	0.199	0.057	0.238	0.777
MLP	train	25.6	0.256	0.195	0.056	0.236	0.923
	test	24.8	0.248	0.196	0.056	0.237	0.830
MS	train	67.6	0.676	0.128	0.024	0.156	0.624
	test	29.0	0.290	0.191	0.053	0.230	0.737

While the ML models were fine-tuned using an extensive range of hyperparameters and various feature selection and engineering methods, they were unable to deliver desirable outcomes. Therefore, we employed classification techniques with discrete target values to observe any improvements in the models' performance.

4.4.2 Evaluating the effectiveness of classification models in predicting success category

We evaluated the effectiveness of the classification models in predicting NIH Percentile categories using the following metrics:

Confusion matrix is a visual representation of actual class vs. predicted class showing information on true positive (TP), true negative (TN), false positive (FP) and false negative (FN) values.

Accuracy score is the percentage of accurate predictions, $(TP + TN)/(TP + FP + FN + FP)$.

Precision is the accuracy of the classifier in predicting TP values, $TP/(TP + FP)$.

Recall or true positive rate is a rate at which the classifier predicts a positive class, $TP/(TP + FN)$.

Specificity or true negative rate is the rate at which the classifier predicts a negative class, $TN/(TN + FP)$.

F1 is an accuracy metric that takes into account both precision and recall, $2(PrecisionRecall)/(Precision + Recall)$.

Classification Error rate at which classifier misclassifies, $(FP + FN)/(TP + FP + FN + FP)$.

Receiver Operator Characteristic (ROC) curve is a graphical representation of the ability of a binary classifier to distinguish classes when the discriminatory threshold is varied. The curve is created by plotting the true positive rate vs. the true negative rate at different thresholds [69].

Area Under the Curve (AUC) is a measure of the ability of a binary classifier to differentiate between classes, with an AUC of 1 indicating that the classifier perfectly distinguishes between classes and an AUC of 0 indicating that the classifier fails at distinguishing classes [69].

The results of evaluating classification ML models are summarised in Table 4.3 as well as in the ROC curves plot (Figure 3.26) and confusion matrices (Figure 3.27). The results clearly show that the XGBC model demonstrates the best predictive power among all classifiers. XGBC model's AUC score of 0.940 is almost the same as the one for the GBC model (0.935), but XGBC has higher accuracy scores and higher F1 score than the GBC model. The KNN classifier has similar accuracy and F1 scores as XGBC, but its AUC is lower (0.882) than the one for XGBC. The XGBC model also has similar or better results for precision, recall, and specificity than all other models. The confusion matrix shows that the true positive class (40.89% of all records) and true negative class (46.45% of all records) were correctly predicted by the XGBC model. False negative and False positive predictions occurred to 8.88% and 3.78% of all records, respectively.

Table 4.3. Comparing Machine Learning Classification Models with 80 % NIH Percentile Cut Off.

Model	Set	Accuracy Score	Precision	Recall	Specificity	F1	Classification Error
LRC	train	0.71	0.71	0.68	0.73	0.70	0.29
	test	0.70	0.71	0.68	0.73	0.70	0.30
KNN	train	0.90	0.85	0.98	0.83	0.91	0.10
	test	0.82	0.76	0.93	0.70	0.84	0.18
RFC	train	0.71	0.75	0.65	0.78	0.70	0.28
	test	0.72	0.74	0.66	0.78	0.70	0.28
GBC	train	0.88	0.93	0.83	0.94	0.88	0.11
	test	0.87	0.92	0.82	0.93	0.86	0.13
XGBC	train	0.92	0.96	0.86	0.96	0.91	0.08
	test	0.88	0.92	0.83	0.93	0.87	0.12

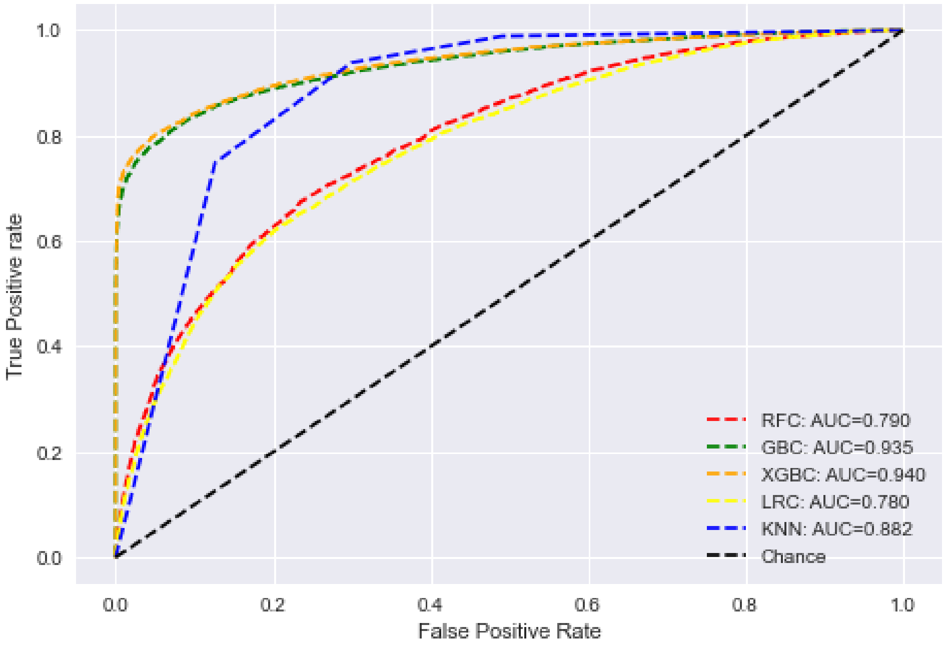


Figure 4.4. ROC curves for classifications models with 80% NIH Percentile cut off.

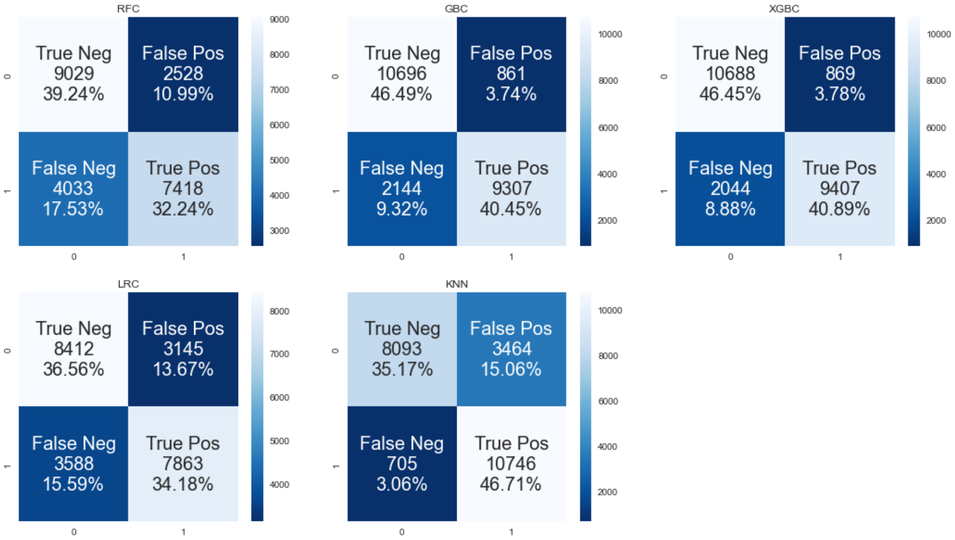


Figure 4.5. Confusion matrices for classifications models with 80% NIH Percentile cut off.

In the study, five models were employed, and it was found that LRC, RFC, and KNN performed relatively poorly compared to the other models. The subpar performance of LRC can be attributed to its simplicity, which may not adequately capture the complex relationship between the target and features in this study. The less favorable performance of KNN could potentially be linked to the large number of features utilized.

Conversely, the two models based on gradient boosting, GBC and XGBC, exhibited the best performance. The superiority of GBC and XGBC over RFC can likely be attributed to the distinct model designs. GBC and XGBC sequentially improve predictions using decision trees, whereas RFC relies on simultaneous predictions from multiple decision trees. Additionally, the slight advantage of XGBC over GBC can be attributed to the regularization technique employed in the XGBC model.

4.4.3 Identifying Important Features

Feature importance determination is an important tool that shows the predictive value of the features used in the model. We applied the Scikit-learn library function $feature\_importances\_$ to the XGBC model [52]. The summary of the feature ranking results presented in Figure 4.6 shows that the Journal Rank is the most influential feature (24.58%) and the Gender of the first author is the least important (2.31%) in predicting the success of the publication. As expected, the year of publication emerged as the second most significant feature. It is reasonable to infer that a long time since the publication date increases the likelihood of accruing additional citations. The

Journal feature ranked third with 9.65% suggesting that the Journal Rank holds greater importance compared to the specific journals where the majority of articles are published in the field of cardiovascular research as shown in section 3.3 and specifically in Table 3.1. It is worth noting that the Institution feature did not show much predictive importance with 4.87%.

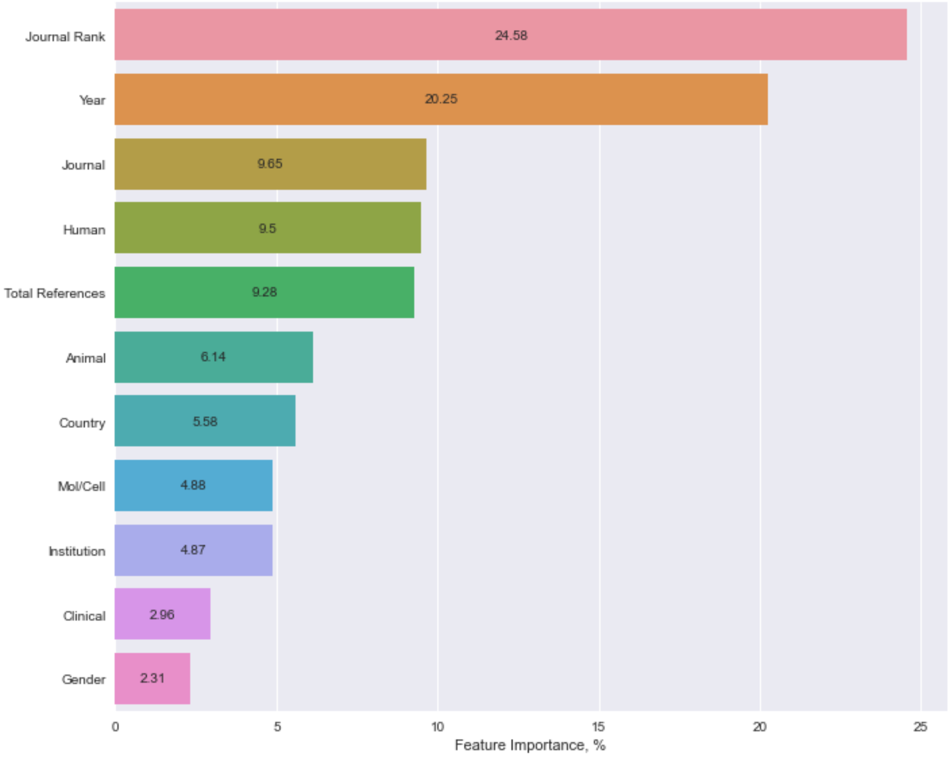


Figure 4.6. Feature importance determined by XGBC model.

CHAPTER 5

Conclusions

This study aimed to identify key factors contributing to successful outcomes in heart disease research. One significant factor examined was the funding provided by the NIH, which showed a positive linear correlation (Pearson coefficient of 0.72) with the number of publications. Additionally, spending caps on scientific research enacted by the US Congress in 2011 is associated with research productivity, as evidenced by a lower annual publication increase rate of 0.7% after 2013 compared to the 3.4% rate observed in the 12 preceding years.

Exploratory data analysis revealed that 21.7% of all heart disease research articles are published in only 10 journals. Interestingly, when considering all journals in the dataset, the journal ranking score according to the SCImago Journal & Country Rank database did not significantly contribute to a higher number of publications. However, when focusing solely on journals in the field of cardiology and cardiovascular medicine, the ranking score became an important factor.

Furthermore, the analysis showed that 16.9% of all heart disease research articles were authored by individuals affiliated with just 10 institutions, indicating a concentration of research output among these institutions.

The study examined the gender distribution in the top 10 journals and institutions for heart disease research. Our findings indicated a higher percentage of male first authors in both the top 10 journals and institutions. Moreover, the analysis revealed that none of the top 10 journals or institutions had a greater number of articles published by female authors compared to male authors.

The study revealed gender disparities in citations and their relationship to the gender of the first authors. Publications with male first authors received double the number of citations compared to those with female first authors. The mean NIH Percentile for papers written by male authors was 1.8 times higher than that for papers with female first authors. Furthermore, when examining the top 10 journals and research institutions ranked by total citations, articles written by male authors had a higher percentage of total citations. The rankings of both journals and institutions varied slightly when considering the gender of the article's first author. These findings highlight the existence of gender differences in citation impact and recognition within the field of heart disease research.

The study analyzed 18 years of publication data from 2002 to 2020 and found a consistent gender disparity in publications within heart disease research. Publications with male first authors outnumbered those with female first authors by a factor of 1.7 throughout the entire period. Although the number of publications increased for both genders over time, there were never more publications written by women than by men. The gender gap remained relatively steady, with 1,180 more publications by male first authors in 2002 and 1,041 more publications in 2020.

The gender gap was also evident when considering the journal ranking based on SJR score. Among the 2,505 journals examined, half as many journals published more articles written by male first authors compared to female first authors (1,558 vs. 832). This gender difference was even more pronounced among the top 100 journals for heart disease research. Similarly, among the 4,053 research institutions analyzed, half as many institutions published more articles written by male first authors compared to female first authors (2,756 vs. 1,262). Notably, none of the top 100 research institutions had a higher number of papers published by female researchers as first authors compared to male researchers.

These findings highlight a persistent gender gap in the publication and visibility of research within the field of heart disease. However, it is important to point out that it was beyond the capabilities of this study to estimate the precise number of women working in heart disease research and this has implications for a true estimation of the gender gap in the CVD research. Based on National Science Foundation records, women in the SEH make up just 36% of both research and teaching faculty positions in academia [70]. Additionally, "in 2018, women made up just 40% of full-time basic science, clinical science, and other health science MD-PhD and PhD faculty at U.S. medical schools" [71]. These numbers are just slightly higher than 33.3% of female first authors publishing heart disease research as found by this study, Section 3.6, and possibly suggest that the gender gap in heart disease research found in this study is a representation of a broader gender gap present in the SEH field in academia. Addressing and rectifying these disparities is crucial to promoting gender equality in scientific research.

This study aimed to identify key factors that contribute to successful outcomes in cardiovascular research impact estimation. Initially, regression models were used to predict the number of people benefiting from a publication, but none of the models performed better than a random model. The accuracy scores were low, even for the training data set, and overfitting was observed in some models. To improve the results, classification techniques were employed using discrete target values. Among the classifiers, the XGBC model demonstrated the best predictive power, achieving the highest accuracy score (0.88), F1 score (0.87), precision (0.92), recall (0.83), and specificity (0.93) on the test data compared to other models. The analysis of feature importance revealed that Journal Rank had the highest predictive value (24.58%), while the gender of the first author had the least importance (2.31%). The year of publication and the journal of publication were also identified as significant factors. However, the Institution feature had limited predictive importance (4.87%). The study found that models based on gradient boosting (GBC and XGBC) performed better than logistic regression (LRC), random forest classifier (RFC), and k-nearest neighbors (KNN) due to their distinct model designs.

BIBLIOGRAPHY

[1] Global Health Data Exchange, “Global Burden of Disease Study 2019 (GBD 2019) Data Resources,” ghdx.healthdata.org. Accessed: Oct. 21, 2021. [Online]. Available: https://ghdx.healthdata.org/gbd-2019

[2] C. J. L. M. Theo Stephen, “Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: A systematic analysis for the global burden of disease study 2019,” *Lancet (London, England)*, vol. 396, pp. 1204 – 1222, 2020.

[3] Wikipedia, “Disability-adjusted life year,” en.wikipedia.org. Accessed: Oct. 21, 2021. [Online]. Available: https://en.wikipedia.org/wiki/Disability-adjusted_life_year

[4] C. P. Gross, G. F. Anderson, and N. R. Powe, “The relation between funding by the national institutes of health and the burden of disease.” *The New England journal of medicine*, vol. 340 24, pp. 1881–7, 1999.

[5] N. Chapman, E. E. Thomas, J. T. Tan, S. C. Inglis, J. H. Y. Wu, R. E. Climie, D. S. Picone, L. C. Blekkenhorst, S. G. Wise, K. M. M. Colafella, A. C. Calkin, and F. Z. Marques, “A roadmap of strategies to support cardiovascular researchers: From policy to practice,” *Nature Reviews Cardiology*, vol. 19, pp. 765–777, 2022.

[6] M. Nicholls, “Funding of cardiovascular research in the usa: Robert califf and peter libby - speak about cardiovascular research funding in the united states and what the latest trends are with mark nicholls.” *European heart journal*, vol. 39 40, pp. 3629–3631, 2018.

[7] John Staddon, “Science and Its Discontents: Too Few Jobs—or Too Many Scientists?” jamesgmartin.center. Accessed: Feb. 20, 2023. [Online]. Available: https://www.jamesgmartin.center/2018/02/science-discontents-jobs-many-scientists/

[8] D. Cyranoski, N. Gilbert, H. Ledford, A. Nayar, and M. Yahia, “Education: The phd factory,” *Nature*, vol. 472, pp. 276–279, 2011.

[9] National Institute of Health, “Funding Rates,” nih.gov. Accessed: Feb. 20, 2023. [Online]. Available: https://report.nih.gov/nihdatabook/category/22

[10] C. A. Chapman, J. C. Bicca-Marques, S. Calvignac-Spencer, P. Fan, P. J. Fashing, J. F. Gogarten, S. Guo, C. A. Hemingway, F. H. Leendertz, B. Li, I. Matsuda, R. Hou, J. C. Serio-Silva, and N. C. Stenseth, “Games academics play and their consequences: How authorship, h-index and journal impact factors are shaping the future of academia,” *Proceedings of the Royal Society B: Biological Sciences*, vol. 286, 2019.

[11] B. Alberts, “Impact factor distortions,” *Science*, vol. 340, pp. 787 – 787, 2013.

[12] M. A. Edwards and S. Roy, "Academic research in the 21st century: Maintaining scientific integrity in a climate of perverse incentives and hypercompetition," *Environmental Engineering Science*, vol. 34, pp. 51 – 61, 2017.

[13] S. Rawat and S. K. Meena, "Publish or perish: Where are we heading?" *Journal of Research in Medical Sciences : The Official Journal of Isfahan University of Medical Sciences*, vol. 19, pp. 87 – 89, 2014.

[14] B. S. Lancho-Barrantes and F. J. Cantu-Ortiz, "Quantifying the publication preferences of leading research universities," *Scientometrics*, vol. 126, pp. 2269 – 2310, 2021.

[15] P. Chatterjee and R. M. Werner, "Gender disparity in citations in high-impact journal articles," *JAMA Network Open*, vol. 4, 2021.

[16] The U.S. National Science Foundation, "The STEM Labor Force of Today: Scientists, Engineers, and Skilled Technical Workers," ncses.nsf.gov. Accessed: Feb. 23, 2023. [Online]. Available: https://ncses.nsf.gov/pubs/nsb20212/participation-of-demographic-groups-in-stem#women-in-stem

[17] Y. A. Shen, J. Webster, Y. Shoda, and I. Fine, "Persistent underrepresentation of women's science in high profile journals," *bioRxiv*, 2018.

[18] K. L. Hart and R. H. Perlis, "Trends in proportion of women as authors of medical journal articles, 2008-2018." *JAMA internal medicine*, 2019.

[19] U.S. National Library of Medicine, "Advanced search results - pubmed," pubmed.ncbi.nlm.nih.gov. Accessed: Apr. 27, 2022. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/advanced/

[20] U.S. National Library of Medicine , "MEDLINE/PubMed data element (field) descriptions," nlm.nih.gov. Accessed: Jul. 07, 2022. [Online]. Available: https://www.nlm.nih.gov/bsd/mms/medlineelements.html#fau

[21] M. Thelwall, "Gender differences in citation impact for 27 fields and six English-speaking countries 1996–2014," *Quantitative Science Studies*, vol. 1, no. 2, pp. 599–617, 06 2020. [Online]. Available: https://doi.org/10.1162/qss_a_00038

[22] SCImago, (n.d.), "SJR — SCImago Journal & Country Rank [Portal]," 2020, scimagojr.com. Accessed: Apr. 27, 2022. [Online]. Available: http://www.scimagojr.com

[23] Institute for Health Metrics and Evaluation, "GBD results," ghdx.healthdata.org. Accessed:Oct. 09, 2021. [Online]. Available: https://ghdx.healthdata.org/gbd-results-tool

[24] U.S. National Library of Medicine, "iCite," icite.od.nih.gov. Accessed: Apr. 27, 2022. [Online]. Available: https://icite.od.nih.gov

[25] Office of Portfolio Analysis, "OPA," dpcpsi.nih.gov. Accessed: Feb. 04, 2023. [Online]. Available: https://dpcpsi.nih.gov/opa

[26] Python, "Python 3.10.1," python.org. Accessed: Oct. 04, 2021. [Online]. Available:

https://www.python.org/downloads/release/python-3101/

[27] Research Organization Registry, “ROR,” ror.org. Accessed: Apr. 27, 2022. [Online]. Available: https://ror.org/

[28] GitHub ROR community, “Affiliation Matching,” github.com. Accessed: Apr. 27, 2022. [Online]. Available: https://github.com/ror-community/ror-api#affiliation-matching

[29] Research Organization Registry, “Retrieve all organization records in ROR,” ror.org. Accessed: Apr. 27, 2022. [Online]. Available: https://ror.readme.io/docs/rest-api#retrieve-all-organization-records-in-ror

[30] Jupyter Lab, “Jupyter Lab,” jupyter.org. Accessed: Sep. 26, 2022. [Online]. Available: https://jupyter.org/

[31] Elsevier, “Scopus,” 2020, scopus.com. Accessed: Jul. 11, 2022. [Online]. Available: https://www.scopus.com/home.uri

[32] V. P. Guerrero-Bote and F. Moya-Anegón, “A further step forward in measuring journals’ scientific prestige: The sjr2 indicator,” *Journal of Informetrics*, vol. 6, no. 4, pp. 674–688, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1751157712000521

[33] Gender-API, “Gender-API,” gender-api.com. Accessed: May. 25, 2022. [Online]. Available: https://gender-api.com

[34] A. Gayet-Ageron, K. B. Messaoud, M. Richards, and S. Schroter, “Female authorship of covid-19 research in manuscripts submitted to 11 biomedical journals: Cross sectional study,” *bmj*, vol. 375, 2021.

[35] L. Santamaría and H. Mihaljević, “Comparison and benchmark of name-to-gender inference services,” *PeerJ Computer Science*, vol. 4, p. e156, 2018.

[36] P. Sebo, “Performance of gender detection tools: A comparative study of name-to-gender inference services,” *Journal of the Medical Library Association: JMLA*, vol. 109, no. 3, p. 414, 2021.

[37] P. Sebo, “How accurate are gender detection tools in predicting the gender for chinese names? a study with 20,000 given names in pinyin format,” *Journal of the Medical Library Association: JMLA*, vol. 110, no. 2, p. 205, 2022.

[38] NamSor, “NamSor-API,” namesorts.com. Accessed: Oct. 03, 2022. [Online]. Available: https://namesorts.com/api/

[39] N. Bérubé, G. Ghiasi, M. Sainte-Marie, and V. Larivière, “Wiki-gendersort: Automatic gender detection using first names in wikipedia,” 2020.

[40] Y. Hu, C. Hu, T. Tran, T. Kasturi, E. Joseph, and M. Gillingham, “What’s in a name?–gender classification of names with character based machine learning models,” *Data Mining and Knowledge Discovery*, vol. 35, no. 4, pp. 1537–1563, 2021.

[41] R. Mehran, A. Kumar, A. Bansal, M. Shariff, M. Gulati, and A. Kalra, “Gender

and disparity in first authorship in cardiology randomized clinical trials," *JAMA Network Open*, vol. 4, no. 3, pp. e211 043–e211 043, 2021.

[42] Pandas, "Pandas," pandas.pydata.org. Accessed: Apr. 15, 2023. [Online]. Available: https://pandas.pydata.org/docs/index.html

[43] S. F. Lee, D. R. Sánchez, M.-J. Sánchez, B. Gelaye, C. L. Chiang, I. O. L. Wong, D. S. T. Cheung, and M. A. L. Fernandez, "Trends in gender of authors of original research in oncology among major medical journals: A retrospective bibliometric study," *BMJ open*, vol. 11, no. 10, p. e046618, 2021.

[44] K. L. Hart and R. H. Perlis, "Trends in proportion of women as authors of medical journal articles, 2008-2018," *JAMA internal medicine*, vol. 179, no. 9, pp. 1285–1287, 2019.

[45] U.S. National Library of Medicine, "PubMed User Guide: Publication Types," pubmed.ncbi.nlm.nih.gov. Accessed: Jul. 19, 2022. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/help/#publication-types

[46] National Institute of Health, "Historical Budget Information," nih.gov. Accessed: Oct. 08, 2021. [Online]. Available: https://www.nih.gov/about-nih/what-we-do/nih-almanac/appropriations-section-1

[47] Wikipedia, "Budget Control Act of 2011," en.wikipedia.org. Accessed: Feb. 14, 2023. [Online]. Available: https://en.wikipedia.org

[48] Ajitesh Kumar, "Correlation Concepts, Matrix & Heatmap using Seaborn," 2022, vitalflux.com. Accessed: Apr. 15, 2023. [Online]. Available: https://vitalflux.com/correlation-heatmap-with-seaborn-pandas/

[49] Pydata, "Seaborn," seaborn.pydata.org. Accessed: Jun. 13, 2023. [Online]. Available: https://seaborn.pydata.org/

[50] Atanu Dan, "Kurtosis() I& Skew() Function In Pandas," 2020, medium.com. Accessed: Apr. 12, 2023. [Online]. Available: https://medium.com/@atanudan/kurtosis-skew-function-in-pandas-aa63d72e20de

[51] H. Kang, "The prevention and handling of the missing data," *Korean journal of anesthesiology*, vol. 64, no. 5, pp. 402–406, 2013.

[52] Scikit-learn, "Scikit-learn," 2023, scikit-learn.org. Accessed: Apr. 13, 2023. [Online]. Available: https://scikit-learn.org

[53] Great Learning Team, "Random forest Algorithm in Machine learning: An Overview," 2023, mygreatlearning.com. Accessed: May. 04, 2023. [Online]. Available: https://www.mygreatlearning.com/blog/random-forest-algorithm/

[54] Arpan Srivastava, "Let's Talk about Random Forests!" 2021, medium.com. Accessed: May. 05, 2023. [Online]. Available: https://medium.com/analytics-vidhya/lets-talk-about-random-forests-524ae1138d8b

[55] Soner Yıldırım, "Gradient Boosted Decision Trees-Explained," 2020, towardsdatascience.com. Accessed: May. 05, 2023. [Online]. Available: https://

towardsdatascience.com/gradient-boosted-decision-trees-explained-9259bd8205af

[56] Albert Um, “L1, L2 Regularization in XGBoost Regression,” 2021, albertum.medium.com. Accessed: May. 05, 2023. [Online]. Available: https://albertum.medium.com/l1-l2-regularization-in-xgboost-regression-7b2db08a59e0

[57] Neetika Khandelwal, “A Brief Introduction to XGBoost,” 2020, towardsdatascience.com. Accessed: May. 05, 2023. [Online]. Available: https://towardsdatascience.com/a-brief-introduction-to-xgboost-3eaee2e3e5d6

[58] R. Rodríguez-Pérez and J. Bajorath, “Evolution of support vector machine and regression modeling in chemoinformatics and drug discovery,” *Journal of Computer-Aided Molecular Design*, vol. 36, no. 5, pp. 355–362, 2022.

[59] Jason Brownlee, “Crash Course on Multi-Layer Perceptron Neural Networks,” 2016, machinelearningmastery.com. Accessed: May. 04, 2023. [Online]. Available: https://machinelearningmastery.com/neural-networks-crash-course/

[60] Sonal Meenu Singh, “What is Multilayer Perceptron (MLP) Neural Networks?” 2023, shiksha.com. Accessed: May. 04, 2023. [Online]. Available: https://www.shiksha.com/online-courses/articles/understanding-multilayer-perceptron-mlp-neural-networks/

[61] Trevor Pedersen, “How To Use “Model Stacking” To Improve Machine Learning Predictions,” 2021, medium.com. Accessed: May. 04, 2023. [Online]. Available: https://medium.com/geekculture/how-to-use-model-stacking-to-improve-machine-learning-predictions-d113278612d4

[62] XGBoost, “XGBoost Documentation,” xgboost.readthedocs.io. Accessed: Jun. 13, 2023. [Online]. Available: https://xgboost.readthedocs.io/en/stable/

[63] Jason Brownlee , “SMOTE for Imbalanced Classification with Python,” 2021, machinelearningmastery.com. Accessed: Apr. 29, 2023. [Online]. Available: https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/

[64] L. Breiman, “Random forests,” *Machine learning*, vol. 45, pp. 5–32, 2001.

[65] Anshul Saini, “Gradient Boosting Algorithm: A Complete Guide for Beginners,” 2021, analyticsvidhya.com. Accessed: May. 05, 2023. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/09/gradient-boosting-algorithm-a-complete-guide-for-beginners/

[66] Nisha Arya, “Logistic Regression for Classification,” 2022, kdnuggets.com. Accessed: May. 06, 2023. [Online]. Available: https://www.kdnuggets.com/2022/04/logistic-regression-classification.html

[67] Antony Christopher, “K-Nearest Neighbor,” 2021, medium.com. Accessed: May. 06, 2023. [Online]. Available: https://medium.com/swlh/k-nearest-neighbor-ca2593d7a3c4

[68] Anuuz Soni, “Advantages And Disadvantages of KNN,” 2020, medium.com. Accessed: May. 06, 2023. [Online]. Available: https:

//medium.com/@anuuz.soni/advantages-and-disadvantages-of-knn-ee06599b9336

[69] Wikipedia, “Receiver operating characteristic,” en.wikipedia.org. Accessed: Apr. 30, 2023. [Online]. Available: https://en.wikipedia.org/wiki/Receiver_operating_characteristic

[70] The U.S. National Science Foundation, “Academic careers,” ncses.nsf.gov. Accessed: Jun. 14, 2023. [Online]. Available: https://ncses.nsf.gov/pubs/nsf21321/report/academic-careers#representation

[71] AAMC, “2018-2019 The State of Women in Academic Medicine: Exploring Pathways to Equity,” aamc.org. Accessed: Jun. 14, 2023. [Online]. Available: https://www.aamc.org/data-reports/data/2018-2019-state-women-academic-medicine-exploring-pathways-equity