Exploring the Relation between Biomedical Entities and Government Funding

Fang Tan
School of Information Management
Sun Yat-Sen University
Guangzhou Guangdong
China
cathytf@163.com

Siting Yang
School of Information Management
Sun Yat-Sen University
Guangzhou Guangdong
China
524228058@qq.com

Xiaoyan Wu
School of Information Management
Sun Yat-Sen University
Guangzhou Guangdong
China
wxy1954174163@163.com

Jian Xu
School of Information Management
Sun Yat-Sen University
Guangzhou Guangdong
China
issxj@mail.sysu.edu.cn

ABSTRACT

In order to study and analyze the effect of government funding on the promotion of scientific research in the field of medicine and to help the government manage research funds more rationally, this study proposes a framework for analyzing the relationship between entities in the field of medicine and funds. The framework consists of four parts: biomedical abstracts acquisition, NIH funding information acquisition and biomedical entity extraction; Development trend analysis of biomedical entity; Analysis of the most funded entities; Analysis of the relationship between entity research popularity and government funding. The results of preliminary analysis are as follows: the field of genetic research is in a period of rapid development, while the field of species research is in a “flat period”; Disease research catch NIH’s continuous attention; the stimulating effect of government funding on the research popularity is decreasing, which is affected by various factors.

CCS CONCEPTS

• Applied computing → Bioinformatics • Applied computing → Computing in government • Information systems → Information retrieval

KEYWORDS

Biomedical entities, Government funding, Entitymetrics, Evolutionary trend

ACM Reference format:

Fang Tan, Siting Yang, Xiaoyan Wu and Jian Xu. 2020. Exploring the Relation between Biomedical Entities and Government Funding. In *Proceedings of 1st Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE2020). ACM/IEEE, Hubei, Wuhan, P. R. China, 5 pages.* https://doi.org/xx.xxxx/xxxxxxxxxx

https://doi.org/10.1145/1234567890

1 INTRODUCTION

By 2019, the total number of literatures in PubMed, the database of biomedical papers, has reached 29 million [1], and statistically, nearly 1/3 of US patents come directly from federally funded programs [2], meaning that the federal government plays an important role in the development of scientific research. Entitymetrics was originally proposed by Ding et al. [3]. Current research around entities in medicine mainly includes the identification and classification of named entities [4], and the extraction of entity relationships [5], while research on government funding is limited to quantifying the effects of government funds in terms of institutions, patents, employment resolution capacity, etc. [6, 7]. Meanwhile, most of the research on scientific research and funding is limited to the exploration of the relationship between some indicators of research achievements (e.g. quantity and citation) and funding, and lacks a detailed study on the impact of funding on entity level. Therefore, this paper combined PubMed medical database and funding information published by the National Institutes of Health (NIH) to compare the actual research focus and funding focus in the biomedical field from 1988 to 2017. First, the trajectory of the field is mapped from a physical research perspective to understand macro trends; second, the most funded entities are counted, the focuses and tendencies of government funding on biomedical entities are summarized; finally, the specific relationship between biomedical research funding and research popularity is further analyzed, which provides a reference for the government 's choice of funding recipients and funding levels.

2 METHODOLOGY

This paper proposes a framework for analyzing the relationship between biomedical entities and funds, as shown in Figure 1.

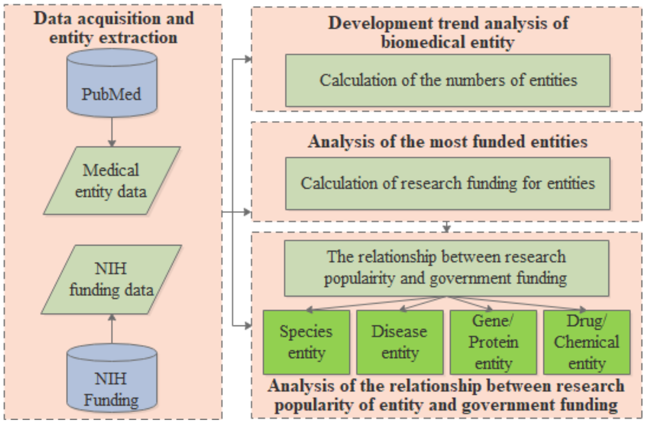

Figure 1. Our Research Framework

In Figure 1, the analysis framework can be divided into four main modules: data acquisition and entity extraction; Development trend analysis of biomedical entity; Analysis of the most funded entities; Analysis of the relationship between entity research popularity and government funding.

1. Data acquisition and entity extraction. Obtaining biomedical data from PubMed between 1988 and 2017, obtaining funding information and relevant research papers of project outputs from NIH funding database, and biomedical entity extraction based on BERN [8, 9, 10]. BERN, namely **B**iomedical named **e**ntity **r**ecognition and multi-type **n**ormalization, a Web-based biomedical text mining tool. The process of entity extraction involving two steps: named entity recognition and entity normalization. At last, 489,433 biomedical entities are obtained between 1988 and 2017. 2,082,652 research projects are obtained, with about $1,0261.3 billion [10].

2. Development trend analysis of biomedical entity. Biomedical entities are categorized into Species, Diseases, Gene/Protein, and Drug/Chemical for evolutionary analysis. Table 1 shows the number of entities of four types.

Table 1. The number of entities of four types

Entity Type	Number
Species	84,203
Disease	36,704
Gene/Protein	25,489
Drug/Chemical	134,574

3. Analysis of the most funded entities. Combined with the biomedical entity data, the entities mentioned in the NIH project output articles are extracted to count the amount of funding for the entities. We define the funding for an entity as the sum of the funding for all articles in which the entity appears.

4. Analysis of the relationship between entity research popularity and government funding. We define the entity research popularity as the number of papers in which the entity is occurred. Thus, the annual number of four types of entities is counted according to the year of research paper in which the entity is located. The years 1988, 1998, 2008, and 2017 are selected, with the entity's research popularity as the vertical axis, and the entity's annual funding amount calculated by step 3 as the horizontal axis to create scatter plots.

3 PRELIMINARY RESULTS

3.1 Development trend analysis of biomedical entity

Based on the change of the number of research entities of each type, the development trend of biomedical fields in the past three decades is analyzed. The number of entities studied in each year is the number of biomedical entity types mentioned in all papers published in that year. Figure 2 shows the number of research entities for each type over time. From the perspective of development trend, the number of gene/protein entities is rising the fastest and is in the stage of rapid development. The research on species entities is in the flat stage and is less numerous.

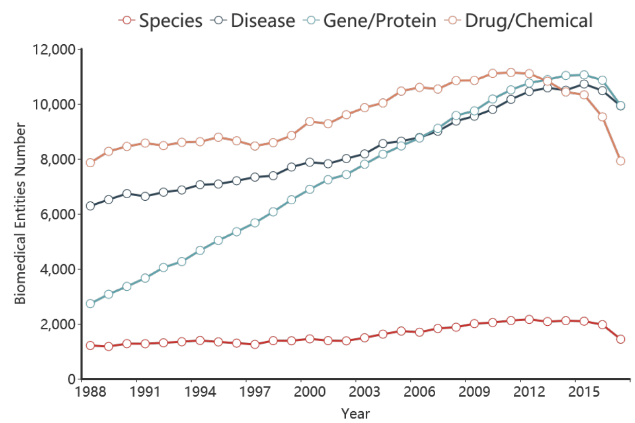

Figure 2. Mention Trends of Biomedical Entities

3.2 Analysis of the Most Funded Entities

Table 2 shows the top twenty biomedical entities in terms of total NIH funding dollars. Mice, HIV, Human immunodeficiency disease and Tumor have all received more than $100 billion, which are the four entities receiving the most funding. In terms of entity types, the highest number of disease entities occupying nine seats, and the lowest number of gene/protein entities, only two. This indicates that the study of disease is an area of research that the NIH has always valued and continues to focus on.

Table 2. Entities with the highest total funding (top 20)

ID	Entity ID	Entity Name	Entity Type	Funds (billion)
1	1009505	Mice	species	183.87
2	1272105	HIV	species	165.38
3	106985801	Human	disease	127.30

		immune-deficiency disease		
4	256225101	Tumor	disease	101.09
5	255268301	Cancer	disease	94.07
6	1009005	Mouse	species	93.22
7	1011605	Rat	species	66.93
8	4168403	Alcohol	drug/chemi-cal	62.13
9	323759402	Insulin	gene/protein	51.24
10	1167605	HIV-1	species	48.70
11	258006601	DM	disease	42.71
12	325454802	CD4+	gene/protein	40.97
13	291977503	Glucose	drug/chemi-cal	39.99
14	107480901	Breast and epithelial-myoepithelial carcinomas	disease	34.50
15	287734103	Ca2+	drug/chemi-cal	28.60
16	107550501	AD	disease	25.93
17	261400701	Obesity	disease	23.81
18	267406001	Depression	disease	23.80
19	106971701	Bronchial asthma	disease	23.65
20	325464002	p32	gene/protein	22.07

3.3 Analysis of the relationship between research popularity of entity and government funding

Based on biomedical entities in the four fields (Species, disease, gene/protein and drug/chemical), the years 1988, 1998, 2008 and 2017 are selected for scatter plotting, and the relationship between entity's research popularity and government funding is visually analyzed, to identify the driving effect of the fund on research in each field from the entity's perspective.

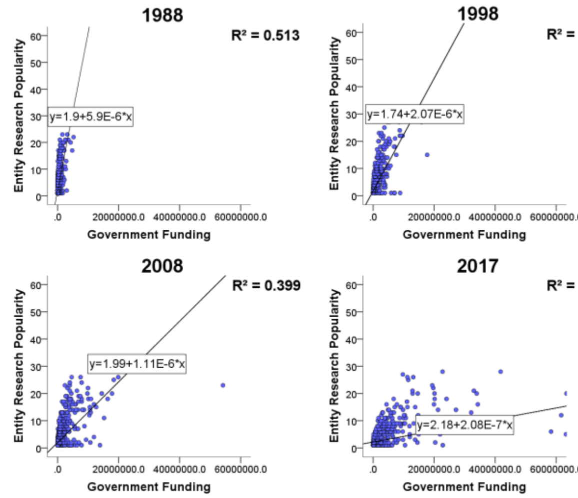

Figure 3. Scatterplot of species entity research funding and research popularity in 1988, 1998, 2008 and 2017

As shown in Figure 3, in 1988, a small increase in funding is followed by a significant increase in research popularity. In the following three years, the linear fit reveals that with the passage of time and the increase of the funding amount, the stimulating effect of funding amounts on the popularity of species research slows down.

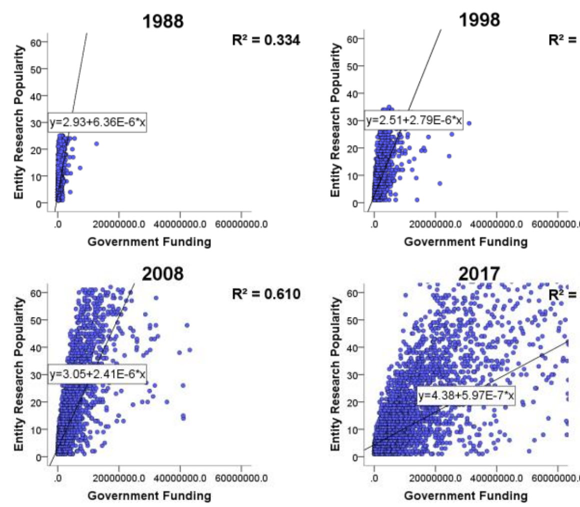

Figure 4. Scatterplot of disease entity research funding and research popularity in 1988, 1998, 2008 and 2017

As shown in Figure 4, The linear coefficient obtained by fitting the linear trend of disease entities in four years is slightly larger than that obtained by species entities. Like the species entity, in 1988, a small increase in funding is followed by a significant increase in research popularity. As the years go by, the increase in funding amount is greater than the increase in research popularity, the slope of the fitted line gradually decreases, which means the stimulating effect of the funding amount on the research popularity is gradually slowing down. In 2017, there are more entities with high funding and low research popularity, which of course could be related to the emergence of new types of entities.

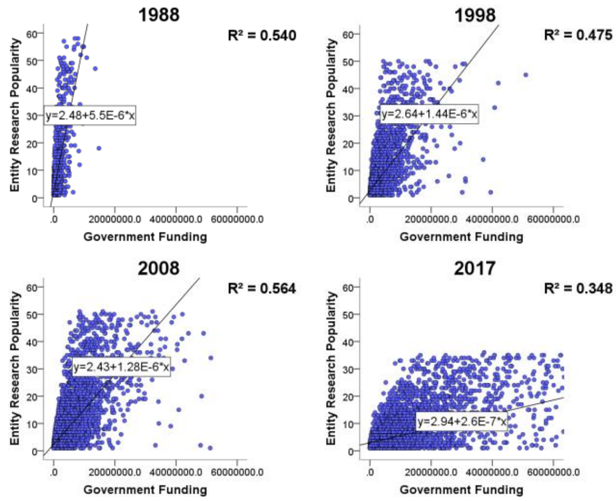

Figure 5. Scatterplot of gene/protein entity research funding and research popularity in 1988, 1998, 2008 and 2017

As shown in Figure 5, for gene/protein entities, the initial trend in 1988 is similar to the first two (species entity and disease entity), while later, especially in 2017, it is clear that the upper limit of research popularity has declined over time.

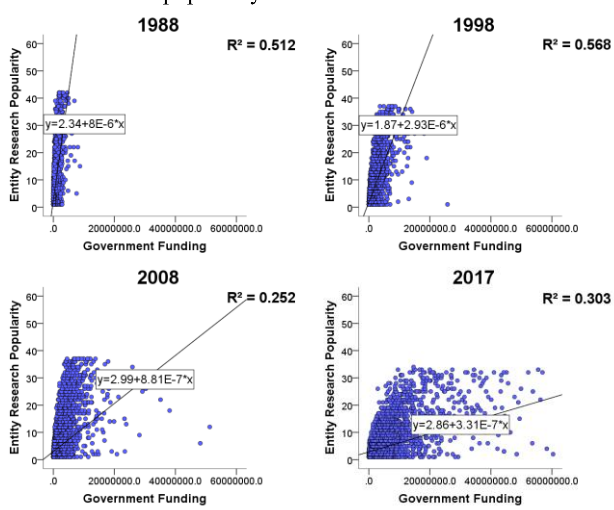

Figure 6. Scatterplot of drug/chemical entity research funding and research popularity in 1988, 1998, 2008 and 2017

As shown in Figure 6, the presentation pattern of drug/chemical entities is similar to that of gene/protein entities. The upper limit of research popularity has declined over time, which indicates that as the years go by, the amount of funding does not play a significant role in the drug/chemical entity's effect on research popularity anymore.

The above analysis follows that the influencing factors of the change of entity research popularity should be multi-faceted and complex, rather than simply being linearly influenced by research funding, and the complexity increases with the increase of years.

4 CONCLUSION AND FUTURE WORK

4.1 Conclusion

Studies linking entities to government funding and exploration of trends from an entity perspective barely visible as far as we know. This study puts forward a preliminary research idea, applying the idea of entitymetrics to biomedical field from the perspective of scientific research funds, and carries out a preliminary research trend exploration and knowledge discovery. The conclusions are as follows: a) the field of genetic research is in a period of rapid development, while the field of species research is in a "flat period"; b) Disease research catch NIH's continuous attention; c) the stimulating effect of government funding on the research popularity is decreasing, which is affected by various factors. These findings provide the basis for a follow-up study.

4.2 Future work

Inspired by the initial results, our future work will focus on a more in-depth exploration of the relationship between government funding and entity development. In this study, we summarized the trends in four categories of entities in the biomedical field and counted the entities that received the highest funding. However, Is there any commonality among these entities? Is entity-related research with any certain characteristics always more likely to be funded by the government? In addition, current research shows that the incentive effect of increased government funding on research in various fields is decreasing, while the impact of other factors such as the continuity of government funding on research popularity has not been explored. Therefore, further research will be conducted on the study of the characteristics of the funded entities and the rules of government funding.

ACKNOWLEDGMENTS

We acknowledge the editors and the anonymous reviewers for insightful suggestions on this work.

REFERENCES

[1] Nicolas Fiorini,,Kathi Canese, Grisha Starchenko, et al., 2018. Best match: new relevance search for PubMed. PLOS Biology 16, 8 (Aug, 2018), e2005343. DOI: https://doi.org/10.1371/journal.pbio.2005343.

[2] Guanghui Xia, Junlian Li, Baokun Xing, et al, 2019. Study of Named Entity Recognition in Medical Treatment Based on Literatures of Chinese Case Reports. Journal of Medical Intelligence 40, 6 (May, 2019), 54-59.

[3] Ying Ding, Min Song, Jia Han, et al., 2013. Entitymetrics: Measuring the impact of entities. PloS one 8, 8 (Aug, 2013), 1-14. DOI: https://doi.org/10.1371/journal.pone.0071416.

[4] Yuan Xu, Yanqiu Ge, Qiang, Wang, et al., 2018. Medical Name Entity Recognition and Application in Chinese Admission Record of Stroke Patients

Based on CRF and RUTA rule. Journal of Sun Yat-sen University (Medical Sciences) 39, 3 (May, 2018), 455-462.

[5] Xiuyan Wang, Lei Cui, 2013. Extract Semantic Relations Between Biomedical Entities Applied Hybrid Method. New Technology of Library and Information Service 29, 3 (Mar, 2013), 77-82.

[6] Yongjian Xu, Jizhong Zhou, 2008. A Study of the Relationship between Government R&D Funding and Business Technology Innovation Activities. China Soft Science, 11 (Nov 2008), 141-148.

[7] Paul A. David, Bronwyn H. Hall, Andrew A. Toole, 2000. Is Public R&D a Complement or a Substitute for Private R&D——A Review of the Econometric Evidence. Research Policy 29, 4-5 (Apr, 2000), 497-529.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, et al., 2018. Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805. Retrieved from https://arxiv.org/abs/1810.04805.

[9] Donghyeon Kim, Jinhyuk Lee, Chan Ho So, et al., 2019. A neural named entity recognition and multi-type normalization tool for biomedical text mining. IEEE Access 7, (Jan 2019), 73729–73740. DOI: https://doi.org/10.1109/ACCESS.2019.2920708.

[10] Jian Xu, Sunkyu Kim, Min Song, et al., 2020. Building a PubMed knowledge graph. Scientific Data 7, 1 (Jun, 2020). 1-15. DOI:10.1038/s41597-020-0543-2.