Medical Research and Innovation: where to look at?

Wei Lu, Todd Rachowin, Yingying Xu, Zhe (Alan) Yu

April, 2017

Section I: Introduction

For our final project, we explore where the best places for health research and innovation are, and how this relates to the rates of specific illnesses. "Where" is meant to be general and refers not only to specific locations, such as cities/states, but also specific diseases, such as obesity, diabetes, and other characteristics. We then rank the 11 top cities for health research and innovation.

Beyond the inherent importance of health research and innovation in treating common and/or life-threatening diseases, substantial resources are spent on them. From 2013 to 2016, the National Institute of Health (NIH) awarded grants totaling over $90B for medical research. Much of the funding goes to major cities, such as Houston, which received $610M from 1,300 grants in 2016 alone. And with the new presidential administration, the NIH might lose over $5B in budget. In addition to public funding, since 2000, over 1000 companies in the medical devices and equipment, healthcare services, and biotechnology sectors received venture capital (VC) financing. All of funding and resources exemplify the importance of understanding where the best places for health research and innovation are.

In order to figure out where the best places are, we first analyze funding on three dimensions: type of disease, time, and location. Based on this high-level analysis, we find location is the most important factor. We deep dive into location, specifically city-level analysis. We breakdown the importance into three categories, which we call: access to talent, access to resources, and access to funding. Based on these three dimensions, we rank the top 11 cities. Based on our ranking system. New York is the top city, while Seattle is the bottom among the 11 cities.

This report is broken down into five remaining sections. Section II describes and summarize the data sources. Section III discusses our high-level analysis and hypotheses based on the type of disease, time, and location. Section IV describes hypotheses and trends in health research and innovation for specific cities. Section V describes our ranking methodology and our final rankings. Section VI concludes the assignment.

Section II: Summary of Data Sources

We use 5 main data sources. The first and major data source that we use is a comprehensive list of all NIH grant recipients, which we gathered directly from the National Institute of Health. The NIH, which is an agency of the United States Department of Health and Human Services, is the primary US medical research agency. It is the largest public funder of biomedical research. Formed in 1887, the NIH is now made up of 27 different components, which are called Institutes and Centers. As of 2016, the agency has a budget of over $31B for medical research that helps lead to breakthroughs and new treatments, as well as help people live longer and healthier lives. This NIH data includes: the grant recipient's organization (e.g. company or university), the recipient's location, and the title of the project that received the grant. The project's title generally includes a high-level description of the recipient's research. We collect data for every 5 years starting from 1995 until 2015 for analyzing the time trends, as well as data from 2012 to 2016 for general analysis.

The other four data sources are: Thomson One's VentureXpert, America's Health Rankings (AHR), the Kaiser Family Foundation (KFF), and U.S. News. VentureXpert contains comprehensive information on VC deals. We look specifically at unique companies that received venture capital funding use data from 2000 to 2016 and are in the medical devices and equipment, healthcare services, and biotechnology sectors (based on PricewaterhouseCoopers MoneyTree). The AHR dataset contains many disease rates in 50 states. We use cancer & cardiovascular death rate and diabetes & obesity rates in each state in 2015. The KFF provides the number of professionally active physicians and hospitals in each state in 2016. The U.S. News dataset contains the top 100 universities (name and location) in the U.S. from 2016.

Section III: High-level analysis of main diseases

In order to understand where health research and innovation is, we first analyze what are the most common illness topics that NIH-funded projects focus on. We believe that NIH will fund diseases that have high incidences and no clear cure, such as cancer, HIV/AIDS. In order to determine the most common diseases that receive NIH funding, we first calculate the frequency of the words used in the projects' titles. We then match the common words to a separate list of medical terms, excluding cancer, that we created. Figure 1 is a word cloud that illustrates the frequencies of projects with the illness terms in 2016.*(Cancer is the highest funded disease. We exclude it since it is double counted when we include the word like "breast" "lung".)

Figure 1: NIH word clound

From the word cloud, it's clear that the most commonly funded projects focused on HIV, brain, and aging (exclude cancer in general). These are similar to what we expected.

Figure 1 provides a baseline for what type of projects NIH provides grants for in 2016, but they it does not explain if this process has changed overtime. For example, it's possible that NIH provides grants for illnesses at specific times based on the incidences of the illness. Figure 2 tracks the funding for projects with the top 10 illness-related terms from 1995 to 2015 for every 5 years.

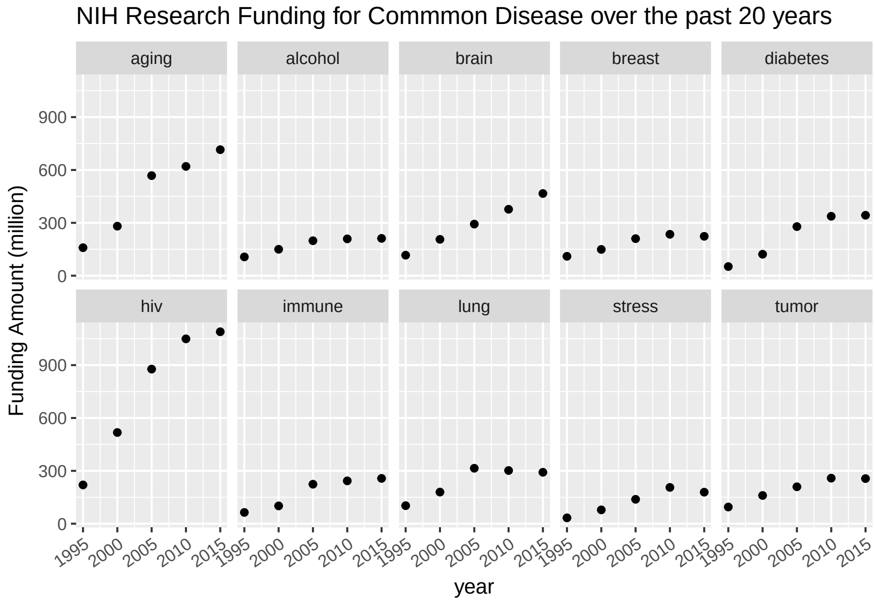


Figure 2: trend for disease funding

For all 10 diseases, there's a clear trend that from years 1995 to 2010, the number of grants per illness-related term increases. This corresponds to higher budgets with NIH, which is as expected. At the same time, there are a few diseases that had higher incidences of funding compared to overall growth. For example, funding for aging and brain increased in 2015 compared to 2010, while NIH's budget slightly decreased during this time. This indicates that there might be specific focuses on diseases in a given year, but there does not appear to be too much leverage.

Next, we analyze specific location for where NIH funding goes. Although some of the results may be obvious when analyzing which states receive funding, such as larger states will likely receive more funding, there may be interesting nuances of why these states receive funding. Figure 3 explores the distribution of funding by state in 2016.

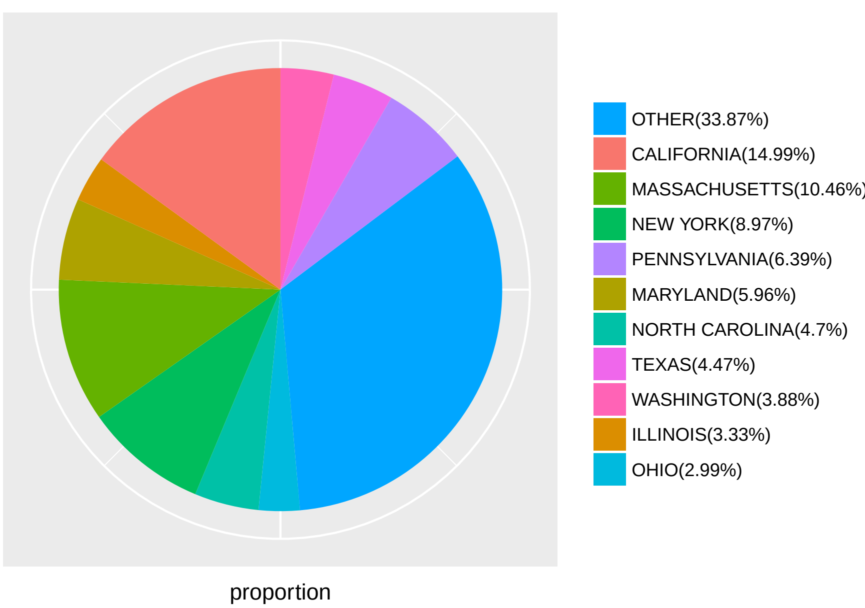
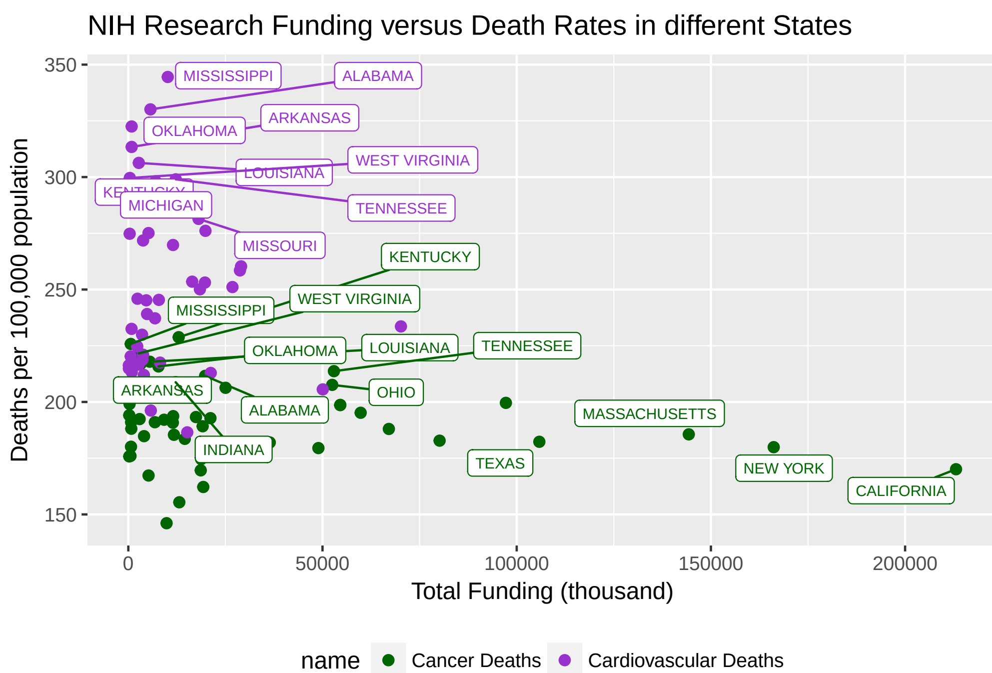


Figure 3: top NIH funding state

As expected, in general, the largest states and those with major cities had the largest amounts of NIH funding regardless of year. California (15%), Massachusetts(10.5%), and New York(9%) have the highest amounts of funding, and the top 10 account for over 66% of total funding, which indicates that substantial amounts of funding is available outside of the major states. Given a clear differentiation between states with large populations, we next explore to see if this state proportion is a result of a relation between instances of diseases. Figure 4 explores the relationship between funding and deaths rates of cancer & cardiovascular reasons and disease rates of diabetes & obesity.



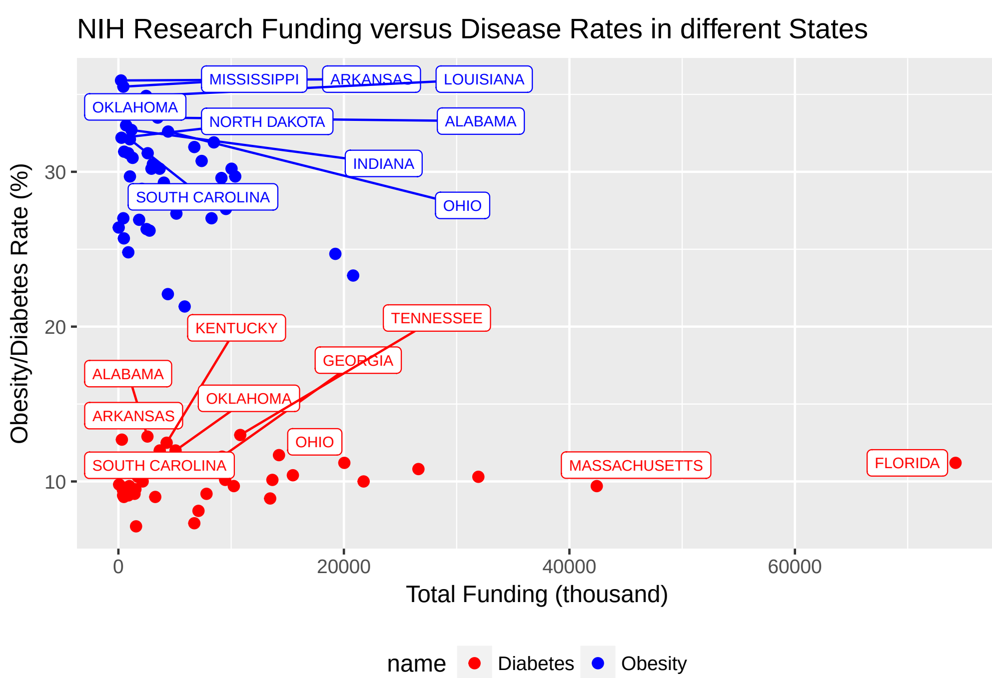


Figure 4: Funding vs Disease

With a lack of a positively sloped linear relation between the variables, the NIH-backed research topics is not highly consistent with the death/disease rate. California and New York have the highest amount of fundings for cancer related research, however, those two states do not show high rates of cancer death. Similarly, Florida and Massachusetts are more curious with diabetes with no significant higher diabetes rates. In addition, the death rate related to cardiovascular disease are even higher than cancer, however, we cannot compare the severity of those two diseases based on the total amount of NIH funding.

We next try to determine if there is any relation between states and specific types of diseases. Figure 5 analyzes the proportion of each state's funding relative to each disease.

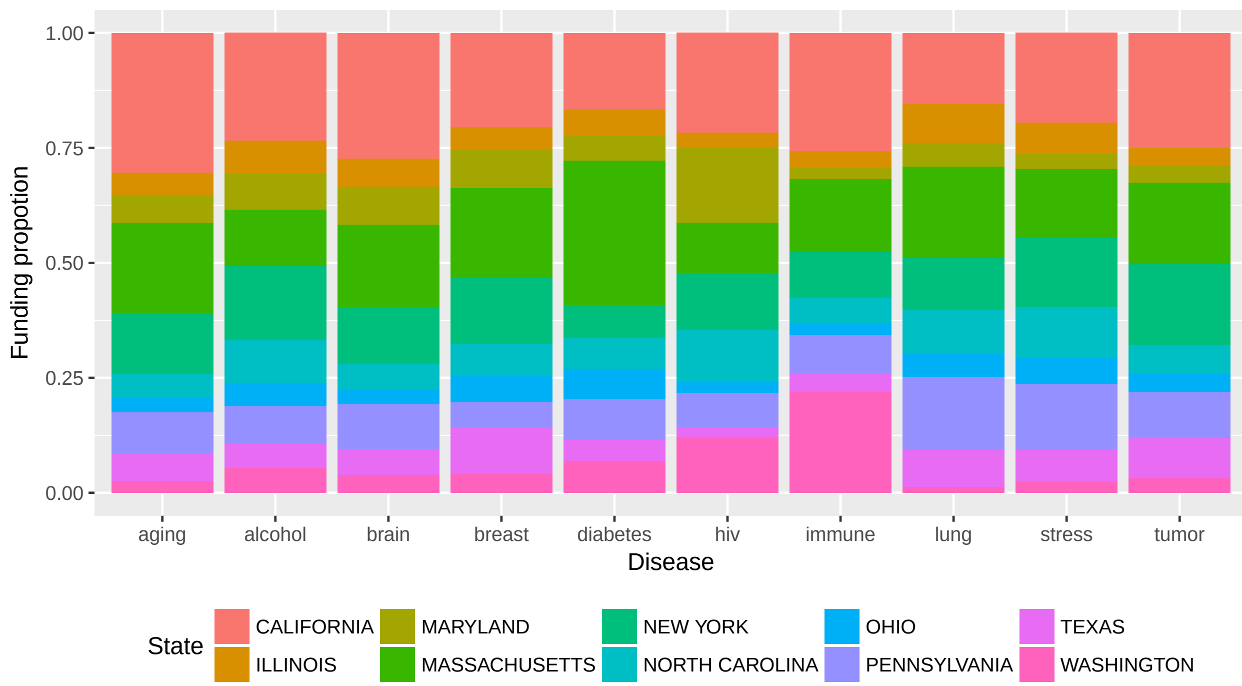


Figure 5: Funding propotion vs Disease

Some states clearly specialize (or not) in specific diseases, such as Washington which has a much higher proportion of relative funding for immune and Massachusetts with diabetes.

With clear trends in medical research and innovation with location, we focus the rest of the analysis on cities and then rank the top 11 cities.

Section IV: Analysis of physical location (cities) of funding

We first want to determine if cities have similar trends to states. Specifically, in Table 1, we look at the top states that received NIH funding and break the states down by city.

Table 1: A City Ranking Table

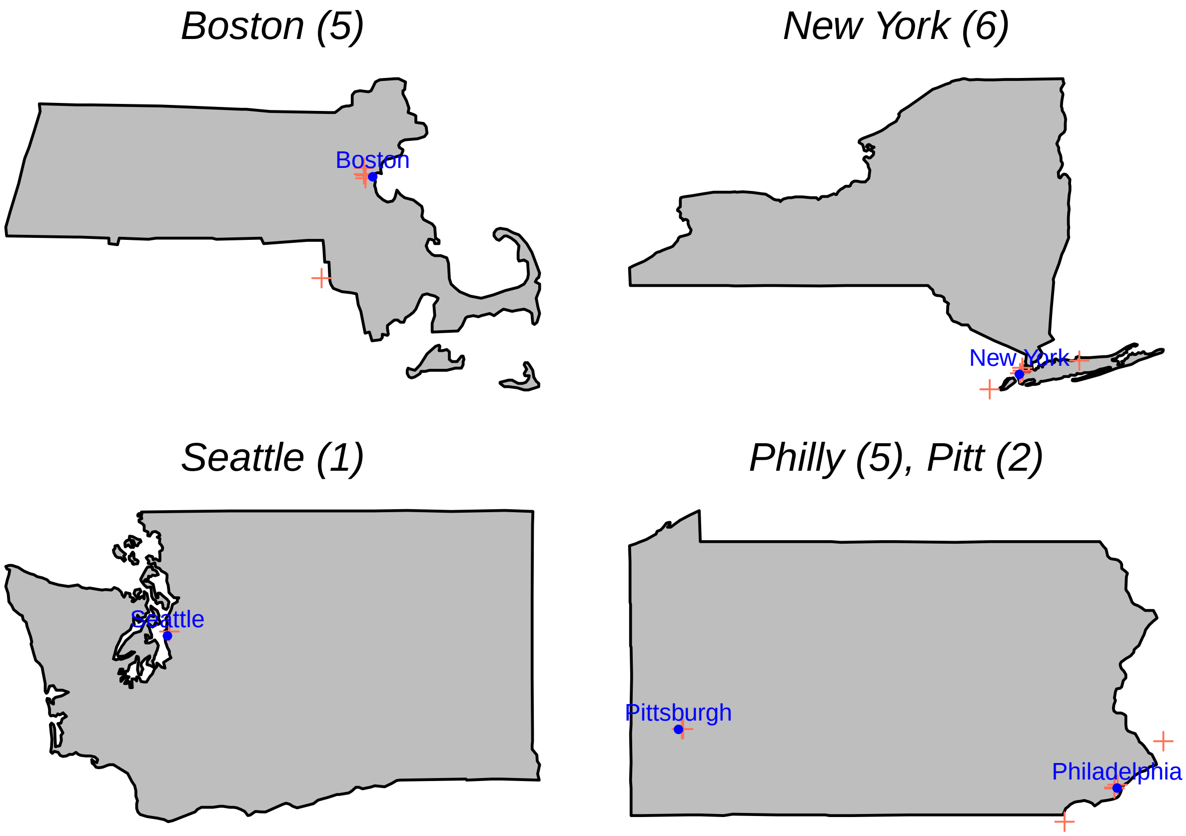
STATE	CITY	Size_State	Funding_State	Pop._US	NIH.Fund_US
Massachusetts	Boston	1	1	24	1
New York	New york	1	1	1	2
Washionton	Seattle	1	1	21	3
Pennsylvania	Philadelphia	1	1	5	4

STATE	CITY	Size_State	Funding_State	Pop._US	NIH.Fund_US
Maryland	Baltimore	1	1	26	5
California	San Diego	2	1	8	6
California	Los Angeles	1	2	2	7
Illinois	Chicago	1	1	3	8
Pennsylvania	Pittsburgh	2	2	62	9
California	San Francisco	4	3	14	10
Texas	Houston	1	1	4	11

We can see that for those largest states, the largest cities receive the most funding. As such, we look at the top 10 cities (note: we combine La Jolla and San Diego) and Houston that received NIH funding for the rest of the analysis.

We next look at three major characteristics for each city: nearby universities (we call access to talent), number of active physicians and hospitals in the state (we call access to resources), and NIH and VC funding (we call access to funding).

Figure 6 shows the number of universities within a Euclidean distance of 1 based on longitude and latitude of each city center. We assume there will not be substantial differences as major cities tend to have a few top universities simply because they are cities.



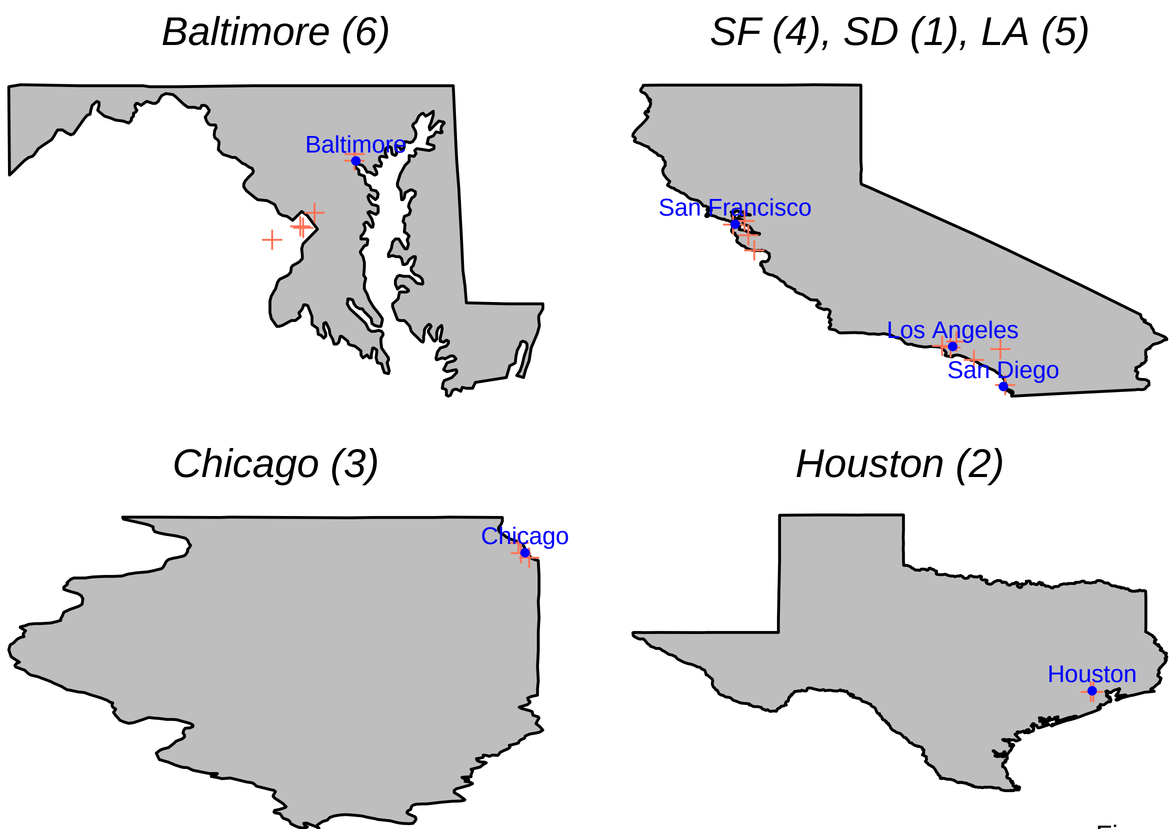


Figure 6

As expected, there is not a significant difference in the number of top 100 universities located in the top cities, but both Boston and New York, 2 of the most NIH funded cities, have two of the highest number of nearby universities. Most of the universities are close to the big cities, but structurally separate and distinct from the big cities, for instance, Texas A&M and Houston.

Figure 7 illustrates the number of physicians and hospitals in the states of the top 11 cities. Similar to NIH funding, we expect the largest states to have the most physicians and hospitals.

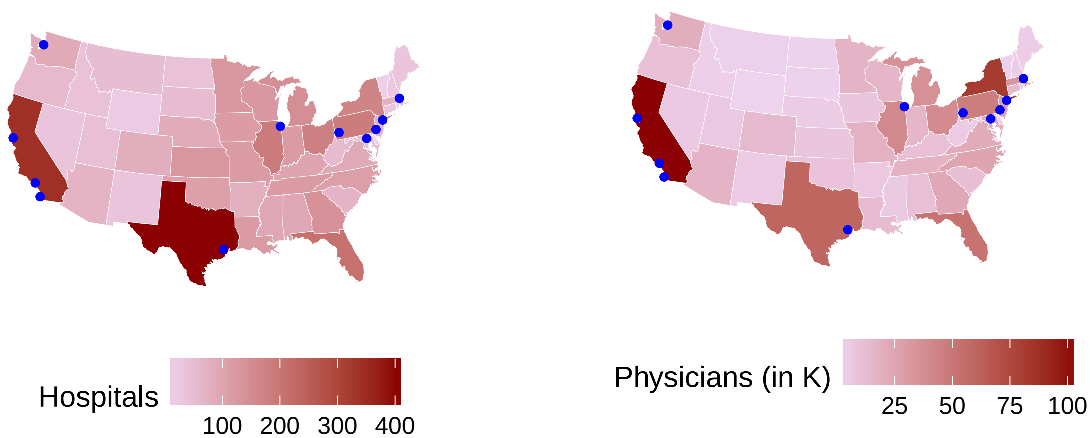


Figure 7: Hospitals and Physicians By State in 2016

Cities in the largest states do in fact have access to the highest number of physicians and hospitals (TX, NY, CA), but that interestingly does not necessarily translate to cities (FL). Texas has over 400 hospitals. California has the highest number of professionally active physicians(>103k), followed with New York (~82k). However, with high the numbers of hospitals and physicians, there is no highly-funded city in Florida.

Lastly, Figure 8 shows the total NIH funding in the past 5 years for the top 11 cities and the number of companies that received VC funding from 2000 to 2016. We expect cities with top universities and high populations to receive the most funding. We also expect the Bay area and Boston to have the most companies that received VC funding because they traditionally had the most VC funding.

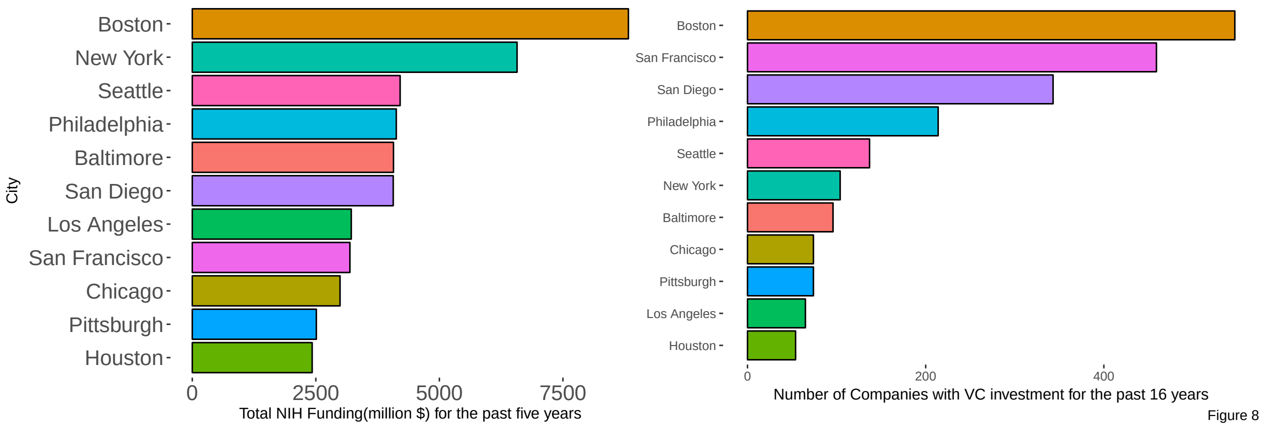


Figure 8

Boston and New York receive a disproportionately high amount of NIH funding. Boston receives 3 times more NIH funding than the No.10 city Pittsburgh. Besides, Boston, SF, and San Diego have a disproportionately high number of companies that receive VC funding.

Section V: Ranking of cities

We rank cities based on the three dimensions previously discussed: access to talent, access to resources, and access to funding. For talent, we proxy this by the number of nearby universities. For resources, we proxy this with state physicians and hospitals. And lastly, we proxy funding by funding from NIH and number of companies that received VC funding.

The last part is the ranking system. We rank those 11 cities under each dimension. Figure 9 shows how Los Angeles is ranked. Here, larger block represents a better ranking.

To create an aggregated rank, we average the rankings within each dimension (e.g. average order of number of physicians and number of hospitals for the access to resources category). We then average across the three dimensions.

Our ranking is as follows:

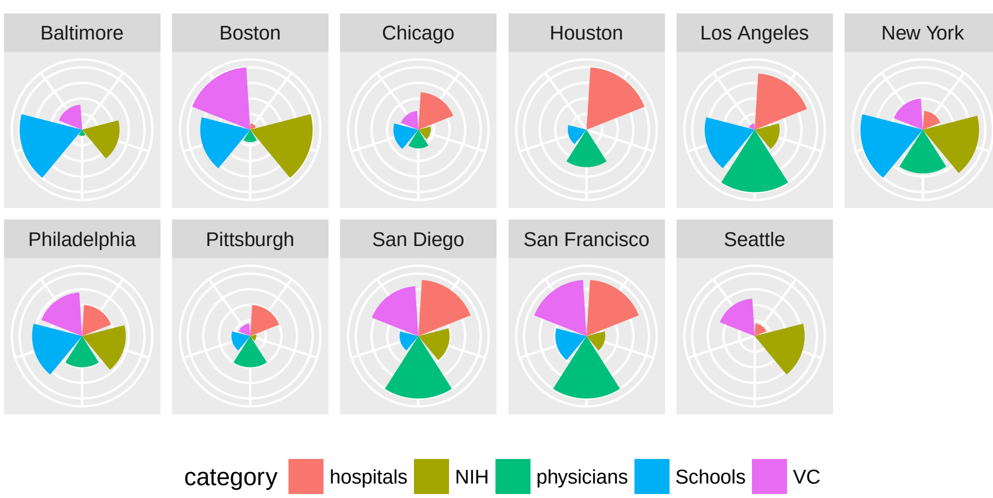


Figure 9

New York is the highest ranked, followed by San Francisco and New York. Seattle is the lowest among the 11 cities. Different cities have their own advantages and strengths, like Boston is the best place in accessing fundings.

Section VI: Conclusion

By analyzing NIH funding, venture capital funding, state active physicians and hospitals, and top universities, we found a few clear results in understanding where are the best places for medical research. For example, the most common illness-related terms are: HIV, aging, and brain (excluding cancer). Also, states seem to receive funding for specific types of diseases that they may specialize in (or do not receive funding for illness terms they do not specialize in).

Based on our ranking system, New York is the best city for medical research and innovation. However, there are several opportunities to extend and improve our analysis. For example, one can specify the question to be related to a specific disease (especially since some states have clear specialties). Additionally, some of the variables can be further specified, such as hospitals near a city versus a state. Whether it's a different question or difficulties finding data, these next steps would provide great extensions and provide an even better high-level understanding of where research and innovation is and this analysis has done so.