Cancer Moonshot: Patent Data Story

Authors

- David Dzamba
- Adam Haiduk (dashboard)
- Zoe Kulsariyeva
- Agata Leszczynska
- Loganathan Ramasamy
- Jakub Smid
- Otakar Smrz (submission)
- Hugo Taborda

Context

The USPTO Cancer Moonshot Challenge invites the public to use patent data to reveal new insights on cancer research and innovative treatments. The challenge is to combine the patent data with other interoperable data sets and present interactive visualizations and stories that can help guide public policy and research to achieve the goal of doubling the rate of progress toward curing cancer or minimizing its impact on lives.

We developed our analysis of the patent data set mostly within the 24-plus hours of the Prague Innovation Days 2016 held by MSD IT Global Innovation Center in Prague.

Approach

We investigated the USPTO Cancer Moonshot Challenge data and discovered relations and implications regarding funding in the various areas of cancer research. Our initial questions were:

1. How effective are cancer research patents with respect to regulatory approvals?
2. What is the landscape of cancer research in terms of patent categories? How does it look for patents with public funding?
3. What is the distribution of different types of cancer in population? How does it compare with public funding and patent efforts?

We explored patent success with respect to FDA approval. We analyzed how NIH funding is distributed across research areas and compared that to cancer incidence data, using NCI data from http://fundedresearch.cancer.gov/nciportfolio/search/SearchForm.

We enriched the patent data set with the full-text contents of these patents downloaded via Google API, where available. We used Python to search for words corresponding to particular cancer types in these patents. We applied the bag-of-words model to represent each patent by the keywords listed in the appendix of the patent data set documentation. We clustered and mapped these keywords into the types of cancer defined in the NCI data to make comparisons possible.

Assumptions

We assume the patent data set contains all records relevant to cancer research. We made particular use of the patent identifier, grant or publication date, CPC-based technology categories, as well as the NIH funding and FDA approval features. However, we did not leverage all of the other promising attributes, such as filing date, patent title, drug or ingredient names, nor the FDA applicant or the NIH grant recipient, as we found those data incomplete or noisy.

For instance, we could not evaluate fully how long it takes on

Dashboard

https://public.tableau.com/views/CancerPatentMoonshot/CancerMoonshotChallenge

Our interactive dashboard is published online at the above link. The key visualizations are exported and attached.

Tableau A

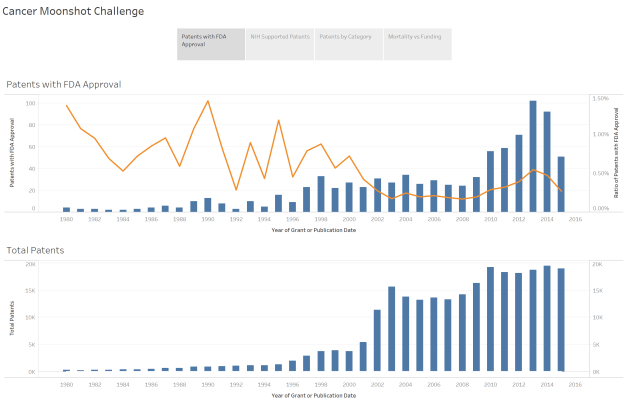


Tableau B

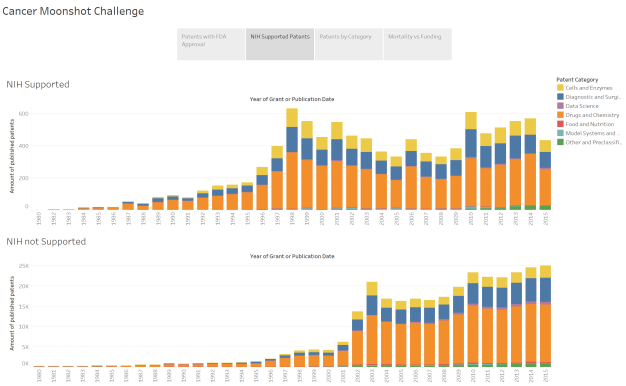


Tableau C

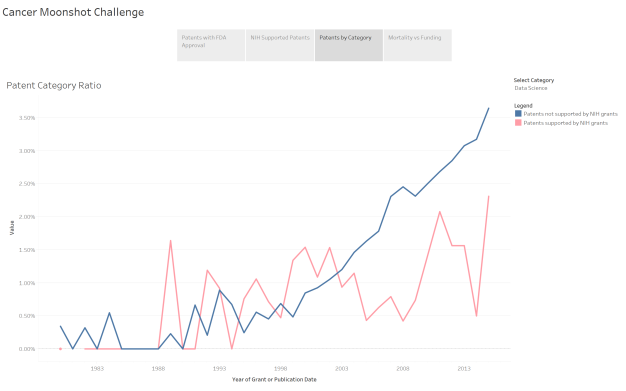
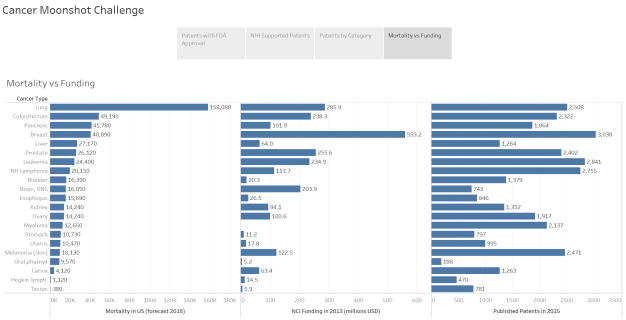


Tableau D

average between a patent is filed and granted or published. The data set only contains patents that were granted or published, and limits those dates to the past 40 years. We could, however, see that for the patents that were granted or published, it took on average about 3 years since their filing, evaluating on yearly basis and not noticing any particular trends.

In our approach and exploration of other interoperable external data sources, we were limited by the 24-plus hours that we could spend on the challenge as a team. We acknowledge this was merely our own deliberate constraint.



Findings

Tableau A explores patent success with respect to FDA approval. The absolute number of published or granted patents is rising around 2002, while the FDA approvals rise only around 2010, with incomplete data since. The ratio of patents with FDA approval shows a deep valley in years 2002-2010. These developments might be associated with the failure of classical drug research around the new millennium and the need for innovation, big pharma mergers and acquisitions, the progress of the Human Genome Project, and later the cancer immunotherapy breakthroughs.

Tableau B and Tableau C analyze how NIH funding is distributed across research areas given the patent data. We found NIH is mainly supporting traditional research like Cells and Enzymes or Diagnostic and Surgical Devices, while an increasing ratio of researchers is focusing on the Data Science area. What other public institution could support data science and research that does not have immediate application in the market?

On the other hand, NIH seems ahead of time in funding cancer-related research that got patented. The industry appears active and ramping up with patents only around the new millennium, whereas the NIH increase comes about 5 years earlier.

Tableau D compares mortality for various cancer types, NCI funding for each of them, and the number of relevant granted or published patents. We enriched the patent data set with the contents of patents via Google API and searched them to derive cancer type relevancy per individual patents. The analysis shows discrepancies in funding and the number of patents for each cancer type when compared with the mortality rate associated with each cancer type.

The Appendix shows the evolution of mortality per various cancer types. Lung cancer shows both increase and decline probably as results of cigarette production and anti-smoking movement.

Conclusion

We have identified historic developments and trends in cancer research using the patent data set. Having enriched that with patent contents via Google API and classified the patents with cancer type, we could compare the distribution of patents to that of NIH funding and cancer incidence in the population.

We suggest that discrepancies in these distributions be taken into consideration by public policy makers and funding institutions. We suggest more funding be directed to research in the cancer types with high mortality yet relatively low funding, like pancreas and liver, and unlike e.g. breast cancer, which seems overfunded.

Appendix

Cancer Facts & Figures 2016. American Cancer Society. Atlanta: American Cancer Society.
http://www.cancer.org/research/cancerfactsstatistics/cancerfactsfigures2016/
http://www.cancer.org/acs/groups/content/@research/documents/document/acspc-047079.pdf

Cancer Statistics, 2016. Siegel, R. L., Miller, K. D., Jemal, A. CA: A Cancer Journal for Clinicians, 66: 7–30.
http://onlinelibrary.wiley.com/doi/10.3322/caac.21332/full
http://onlinelibrary.wiley.com/doi/10.3322/caac.21332/epdf