Investigating the Impact of NLM Resources Using Bibliometric Analysis

Candace Norton, NLM Associate Fellow 2016-2017

August 21, 2017

Project Sponsored by Dr. Dina Demner Fushman, Lister Hill National Center for Biomedical Communications

Contents

Abstract

Objective

The project, Investigating the Impact of NLM Resources Using Bibliometric Analysis, provides an analysis of the uses and potential impact of NLM resources as reported in the published biomedical literature. This report will identify patterns of resource use, and provide a possible pathway to documenting product impact, while showcasing the reach of NLM products to biomedical research communities. The goal of this investigation is to determine if bibliometric analysis can be used to determine product impact.

Methodology

An initial search of each of the 294 NLM databases and APIs, as listed on the NLM main web site (https://wwwcf.nlm.nih.gov/nlm_eresources/eresources/search_database.cfm), was conducted in PubMed using a title/abstract search to identify possible candidates for further research. Five resources were selected for further investigation: ClinVar, Open-i®, OSIRIS, TOXNET, and the Visible Human Project. These resources were selected based on the recommendation of the project sponsor, a knowledgeable expert on NLM resources. Subsequent searches on the identified resources were conducted as title/abstract queries in IEEE Xplore, LISTA, Scopus, and Web of Science. Data extraction related to author affiliation, institution status, geography, and product use was collected from the results of each resource's search results. Based on the total number of citations to review, a random sample of 20% of the results from ClinVar, TOXNET, and the Visible Human project were submitted to full-text review to assess NLM product impact as the relevant criteria was not always available in the bibliographic record. Open-i and OSIRIS were excluded from the full-text review based on too few relevant results.

Results

The initial searches in PubMed revealed that NLM resources are being used by researchers who participate in scholarly publishing. The more focused inquiry of the five targeted resources indicated that researchers from multiple sectors are using NLM products in pursuit of their work with most published research coming from academic, government, and corporate research entities. The most common use of these select NLM products is in developing educational training and in providing a validated reference source for new research. The full-text review of references for ClinVar, TOXNET, and the Visible Human Project revealed that it is possible to determine impact from some references but not all. NLM products have supported the creation of new tools, software, training and education resources.

Discussion

Bibliometric analysis can provide some data that can be used to evaluate product impact: It can answer how many publications have been released citing a product, who is publishing those papers, where they are located, and to some extent how the products are being used and in naming what new tools, methods, or tests have been developed. The limitations of this method of analysis include basing the research exclusively in the published biomedical literature; social media, grey literature, and other resources were not included in the scope of this work. Bibliometric analysis can be used in conjunction with other usage metrics to evaluate the impact of a product.

Background

NLM produces many databases, tools, and products to meet the information needs of a wide and varied user base. Through an ever-growing assortment of metrics, from web analytics to focus groups, NLM can assess the types of institutions, organizations, and individuals who make use of our resources. In many instances, we can even gain an understanding of how products are used and work in partnership to improve the product's usability and enhance the user's experience. It is important to go a step beyond how the users are accessing and utilizing the products and services we provide to better understand and highlight the impact of our resources on their communities of practice. This project is a bibliometric investigation examining how NLM resources are recognized in the published biomedical literature and further exploring the impact of NLM resources in the research output.

Objective

This project investigates a possible avenue for assessing the impact of NLM products through a bibliometric analysis of the published biomedical literature. This report will identify patterns of resource use, and provide a possible pathway to documenting product impact, while showcasing the reach of NLM products to biomedical research communities around the world. The goal of this investigation is to determine if bibliometric analysis can be used to determine product impact, and inform future explorations on product impact assessments.

Literature Review

Bibliometric analysis refers to the practice of applying statistical methods and study to bibliographies (Hood, 2001). This way of evaluating bibliographic data is being used to measure publication activity, subject category trends, geography and co-authoring networks (Thomson Reuters, 2008) with the goal of assessing the value and impact of the research outputs. IMLS and NILPPA encourage libraries to broaden their definition of impact by incorporating indicators of deepening impact, such as the generation of new questions (IMLS, 2017; NLPPA 2014). This project incorporates the foundations of bibliometric analysis and investigation (Broadus, 1987) to identify an isolated product's bibliography, and uses the results of this inquiry to assess the product's impact beyond the initial research output.

Methodology

The National Library of Medicine produces 294 tools, products, and resources that are free and open to the public to use. The first stage of inquiry involved searching each resource by name in PubMed. The specific resource was searched in quotation marks as a phrase, first in the title only, then in both the title and abstract. These searches are included in supplemental document A (SupplA_PubMed Scoping Search NLM Products).

The next stage of the inquiry involved selecting a representative sample of resources to explore in more depth. ClinVar, Open-i®, OSIRIS, TOXNET, and the Visible Human Project were chosen. Selection criteria included identifying a balanced mix of legacy resources with more recently developed products, selecting a range of resource types (bibliographic resources, image resources, and genetics resources), and all products selected had a moderate number of results in the PubMed scoping searches.

Each resource was searched by name as a phrase demarked by quotations in the title and abstract in five databases: PubMed, LISTA, IEEE Xplore, Scopus, and Web of Science. Search strategies and results are included in supplemental document B (SupplB_NLM Products Data Extraction).

Results from each search were exported to an EndNote library, available as supplemental document C (SupplC_EndNote Library NLM Products). Results were pooled across databases for each resource; the “find duplicates” feature of EndNote was used in conjunction with manual screening to remove duplicate references from each pooled set of results. Appendix 1 documents the initial number of results for each search, the number of duplicates removed, and the final number of results screened for each resource.

Data extraction for each set of results was designed to capture data points of interest to NLM: Author affiliation and type of organization, date of publication, country and/or state location, and reported product use. This extensive data extraction was completed for all relevant results for each product. For data extraction related to the impact of the NLM product as reported in the literature, a full-text review was required to locate this data.

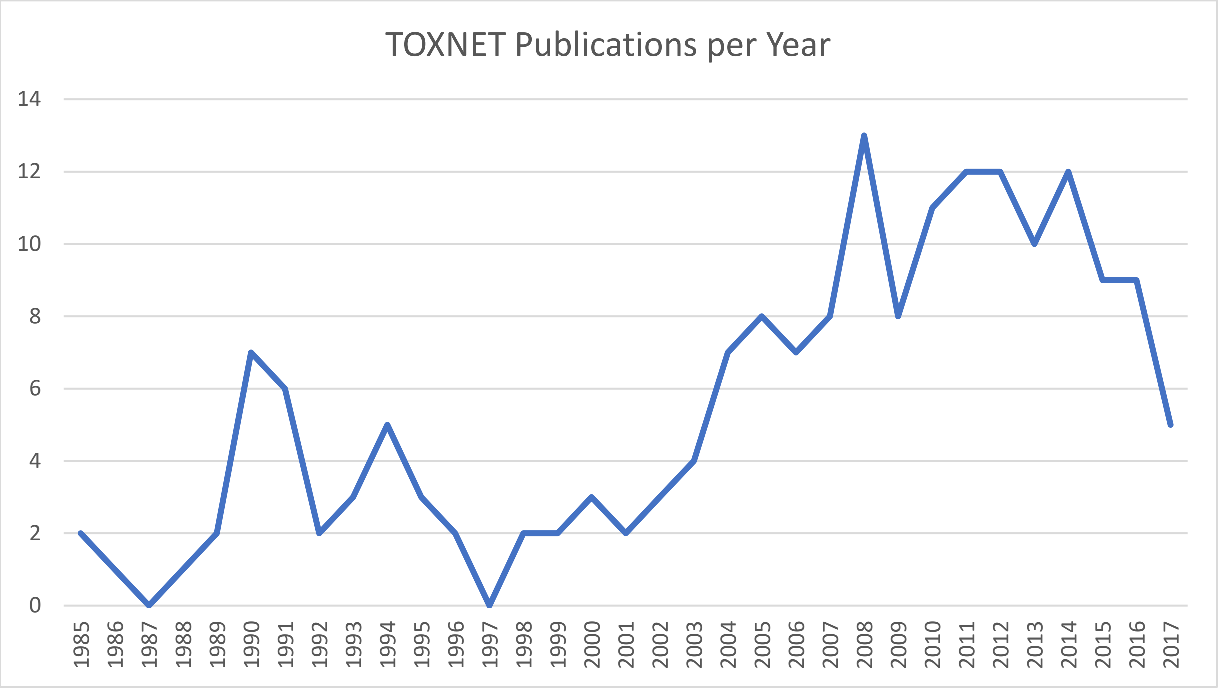
Based on the total number of relevant citations for the project and the associated time frame to complete the work, a random sample of 20% of the relevant results for each ClinVar, TOXNET, and The Visible Human Project was selected for full-text review. Open-i and OSIRIS had too few relevant results to sample. The data extraction tables are available in supplemental document B, and details on the sampling process are included in Appendix 2.

Results

TOXNET

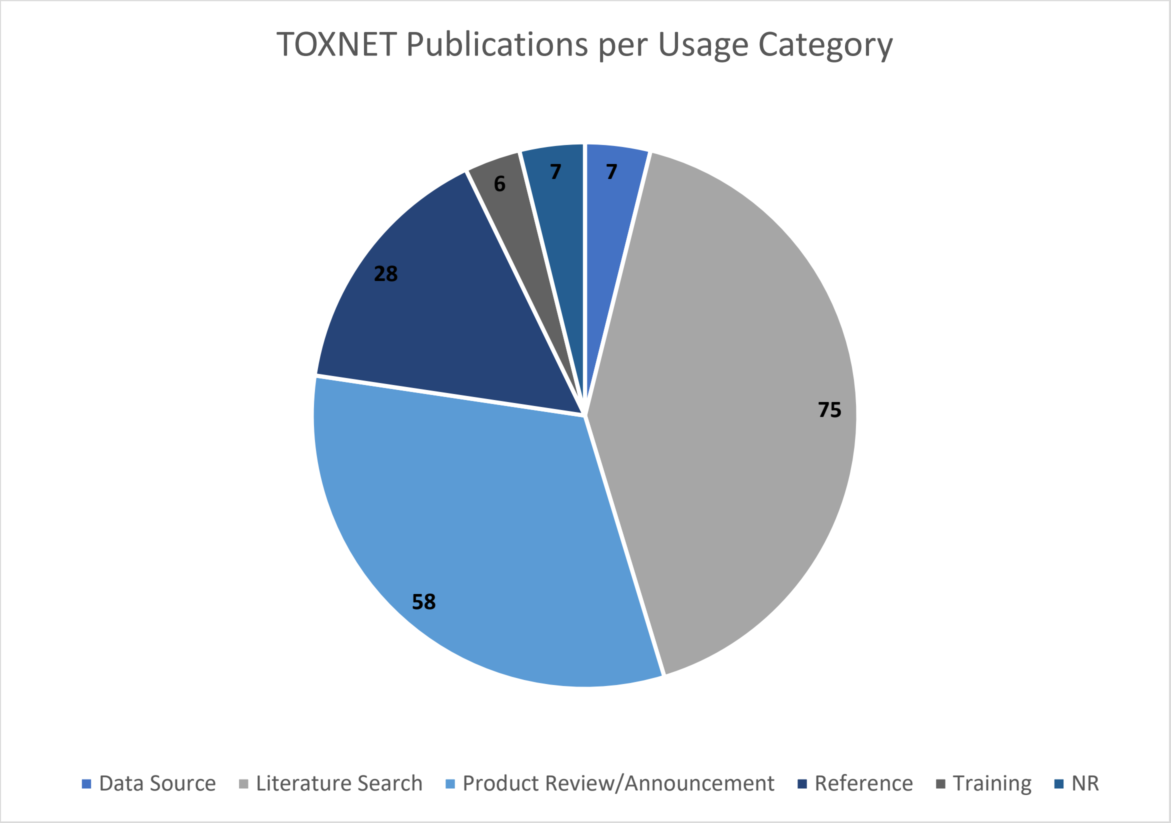
In total, after removing duplicates, the TOXNET review included 181 bibliographic records. TOXNET, established in 1985, saw the most references in published papers in 2008 with 13 publications (7%) naming TOXNET as an information resource. An NLM author was the lead author on 39 of the 181 publications, or 21% of the publications.

Figure 1. TOXNET Publications per Year (1985-2017)



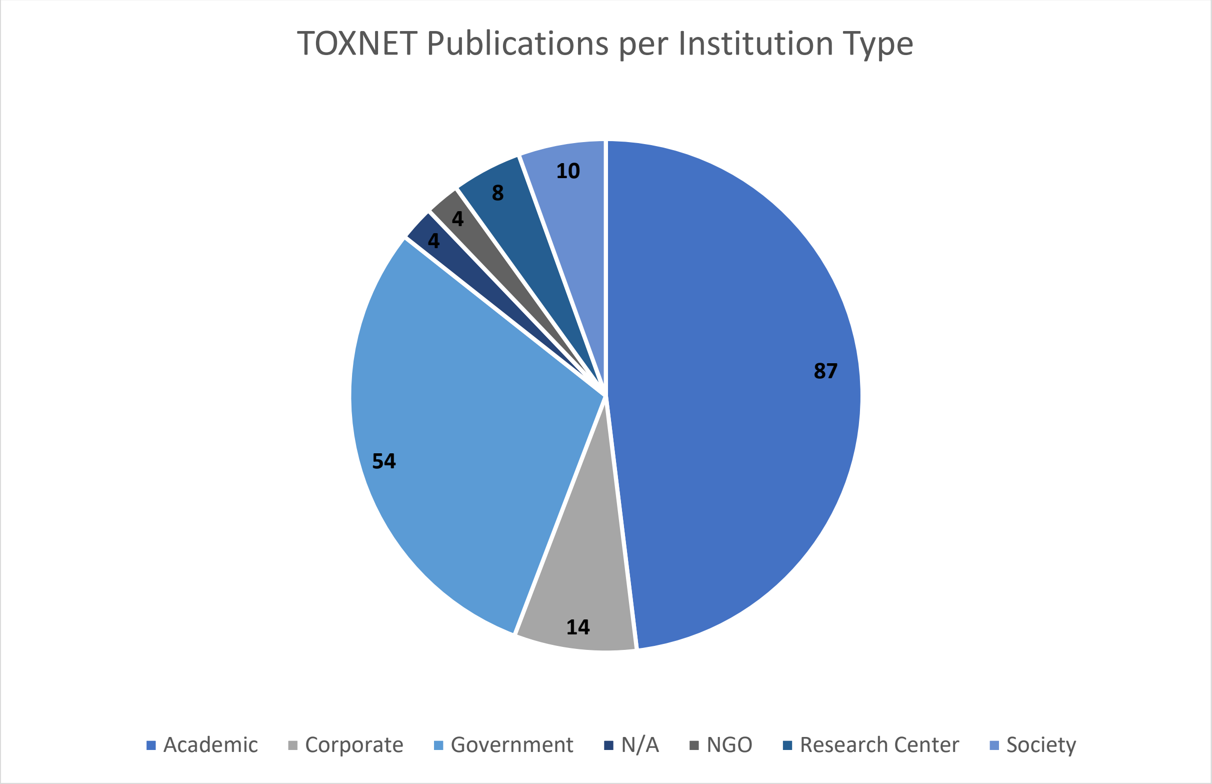
Overwhelmingly, TOXNET was most frequently referenced as a resource searched in a literature review (n=75, or 41%) closely followed by publications that provided an announcement about new content being added to the database or articles that included a product review or introduction (n=58, or 32%).

Figure 2. TOXNET Publications per Usage Category



Most of the publications mentioning use of TOXNET were affiliated with an academic institution (n=87, or 48%), with government organizations producing the next highest volume of papers (n=54, or 30%).

Figure 3. TOXNET Publications per Institution Type



TOXNET has been referenced by researchers from 20 countries and 28 US states. After the US (n=118, or 65%), the countries with the highest number of references are Italy (n=18, or 10%) and Canada (n=13, or 7%). Within the US, the states with the highest number of publications documenting use of TOXNET are Maryland (n=42, or 36% of all US publications and 23% of the total publications) and New York (n=10, or 8% of all US publications and 5% of the total publications).

Figure 4. TOXNET Publications per Country

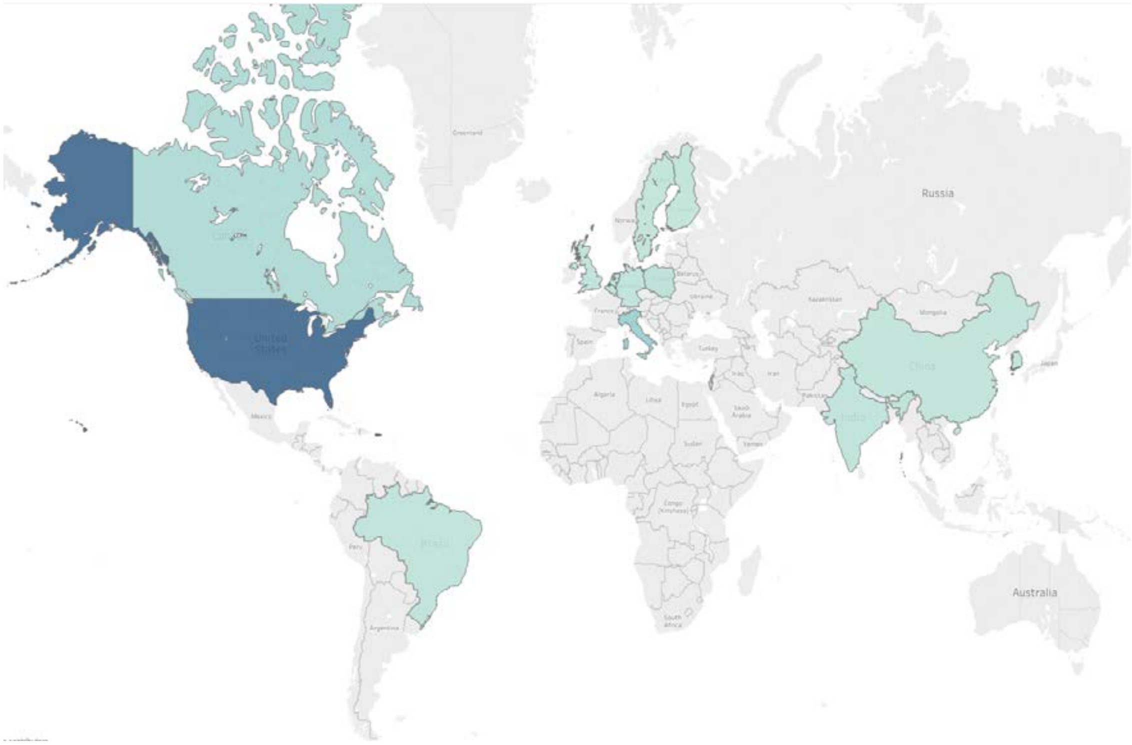
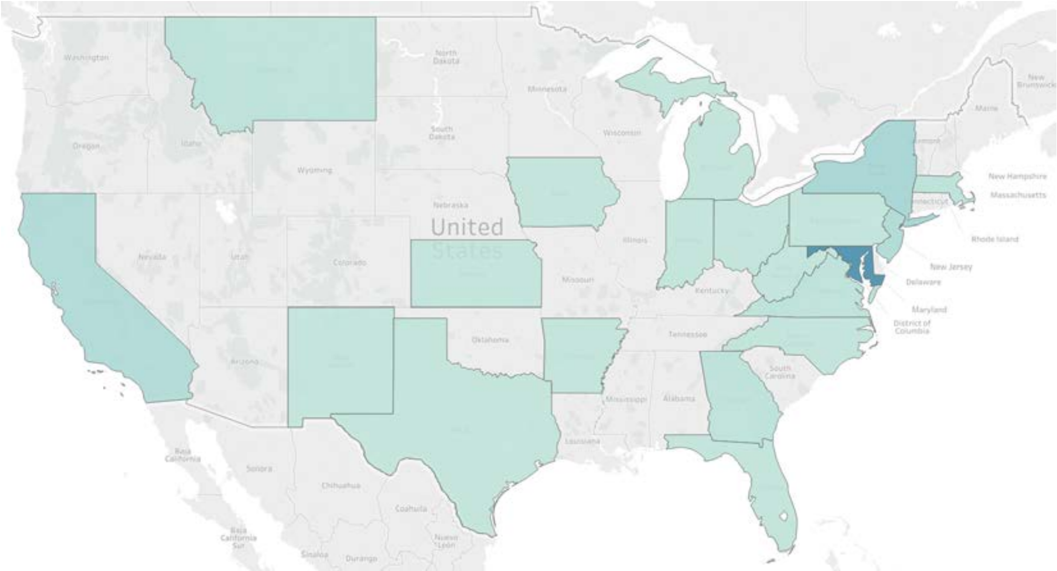


Figure 5. TOXNET Publications per US State

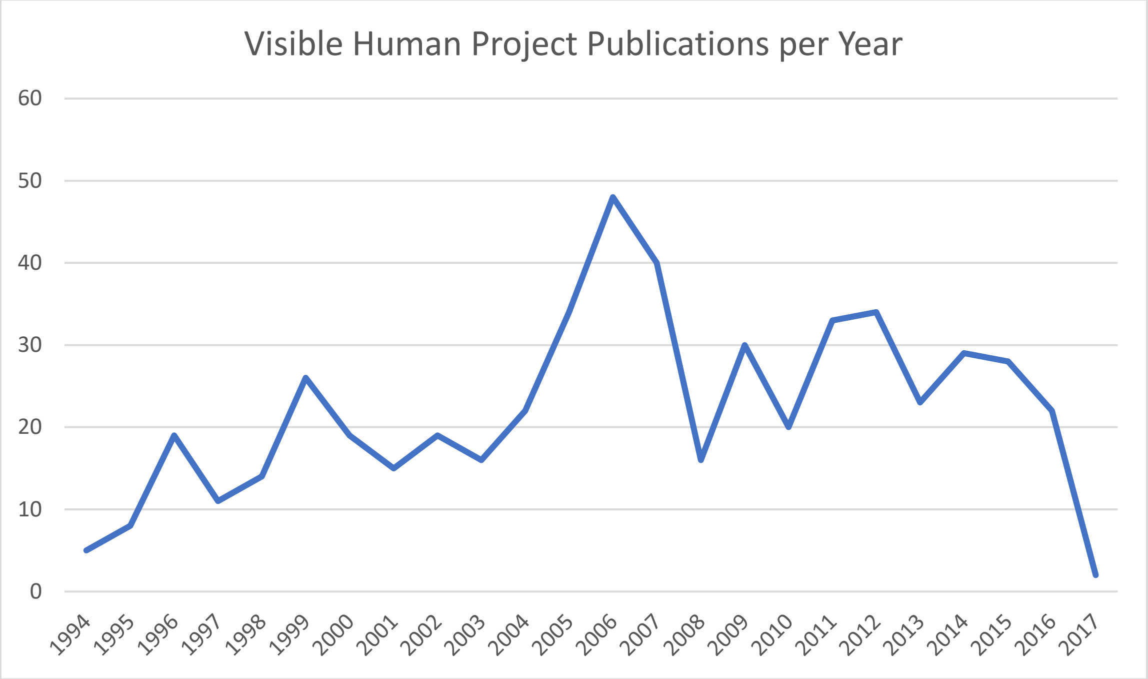


Twenty percent of the total TOXNET results were reviewed in full-text to identify impact indicators as documented by the manuscript authors. Thirty-six papers were reviewed in full-text, with eight papers detailing how using TOXNET improved diagnosis time, provided cross-linked data to other federal databases for more robust research sources, and provided validated information for chemical risk assessment in epidemiological studies among other examples.

Visible Human Project

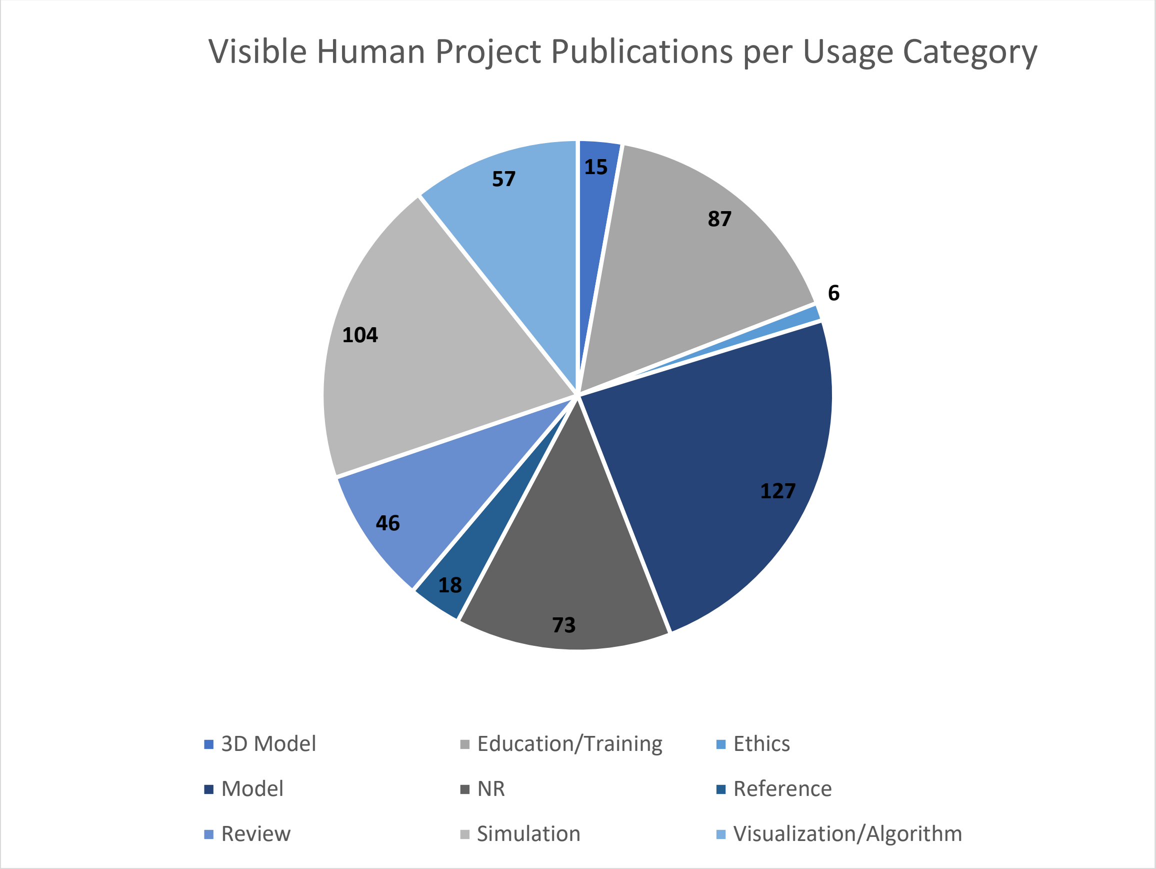
The Visible Human Project had the largest number of references to review, totaling 533 after removing duplicates. The Visible Human Project was proposed in 1989 and the male dataset was completed in 1995, the female dataset in 1994. The Visible Human Project saw the highest number of citations in the literature in 2006 (n=48, or 9%). An NLM author was the lead author on 26 of the 533 publications, or 4% of the publications.

Figure 11. Visible Human Project Publications per Year (1994-2017)



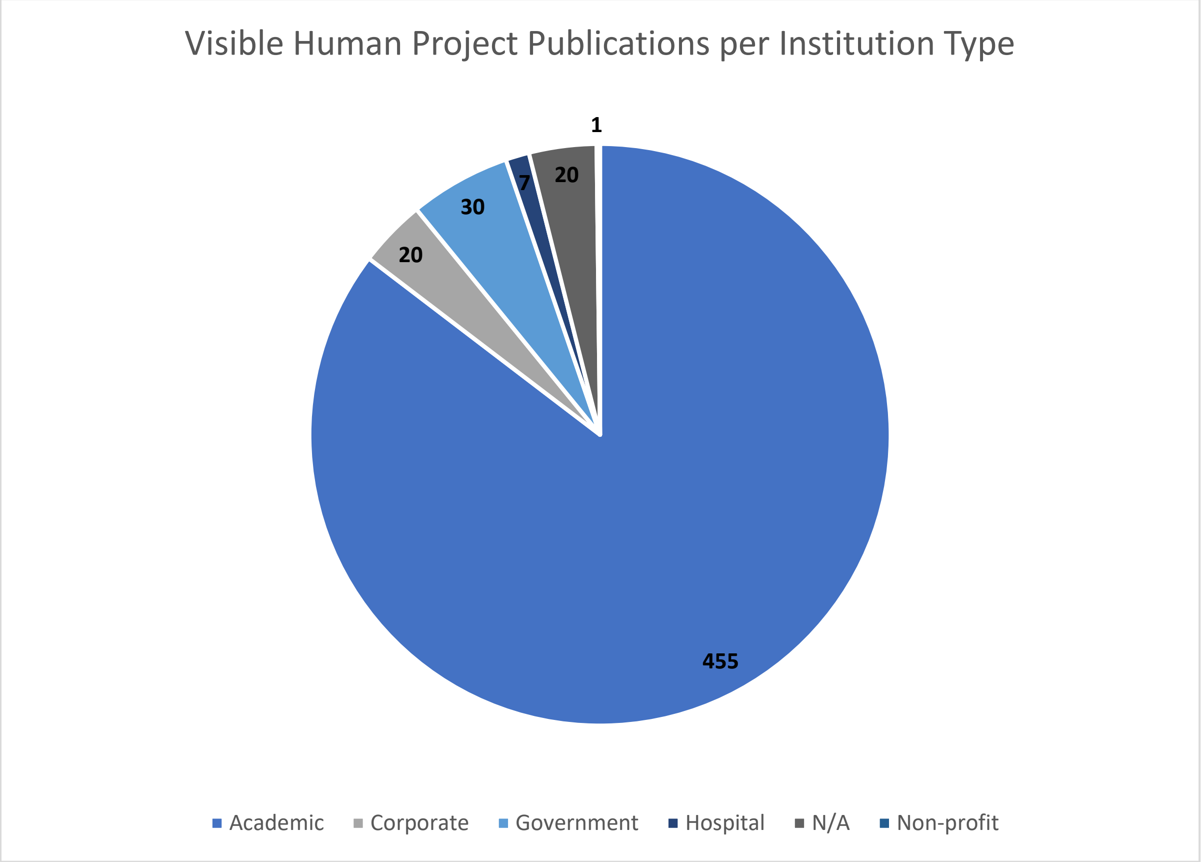
The Visible Human Project had the greatest variety of uses of the three resources investigated in this project. The highest volume of use was noted in reports of using the datasets to create new models (n=127, or 24%), followed by being source material for simulations (n=104, or 20%), and various uses in education and training (n=87, or 16%).

Figure 12. Visible Human Project Publications per Usage Category



Overwhelmingly, researchers who published papers referencing the use of the Visible Human Project were affiliated with an academic institution (n=455, or 85%). The remaining publications were predominantly split between government (n=30, or 6%) and corporate (n=20, or 4%) authors.

Figure 13. Visible Human Project Publications per Institution Type



The Visible Human Project has publications referencing the use of the datasets originating from authors in 41 countries. Outside of the US (n=187, or 35%), the countries with the most publications are China (n=105, or 20%) and Germany (n=32, or 6%). Within the US, the states with the highest number of publications are Maryland (n=28, 15% of all US publications or 5% of the total publications) and New York (n=21, or 11% of all US publications or 4% of the total publications).

Figure 14. Visible Human Project Publications per Country

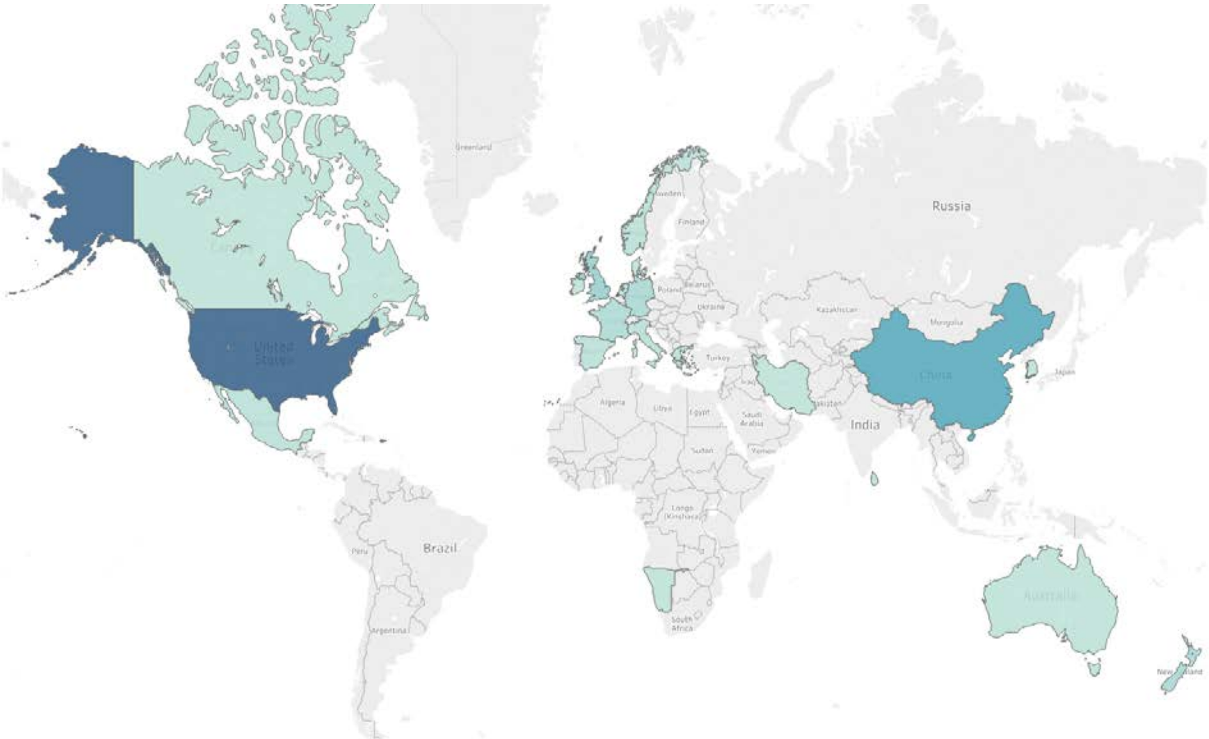
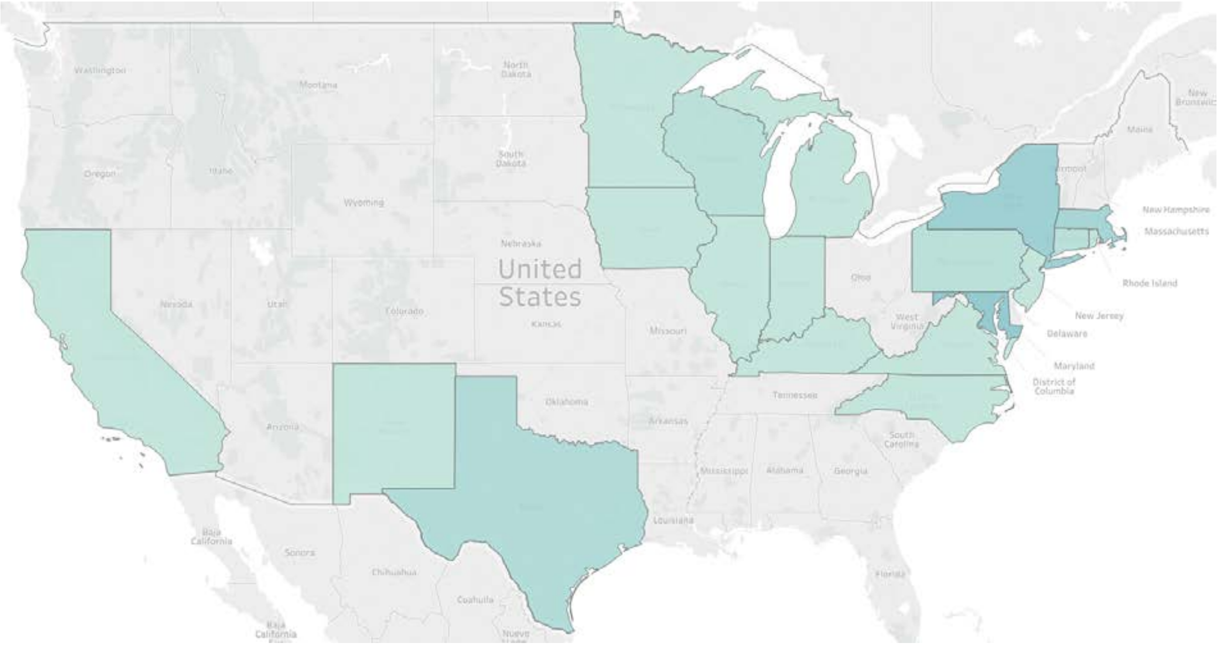


Figure 15. Visible Human Project Publications per US State

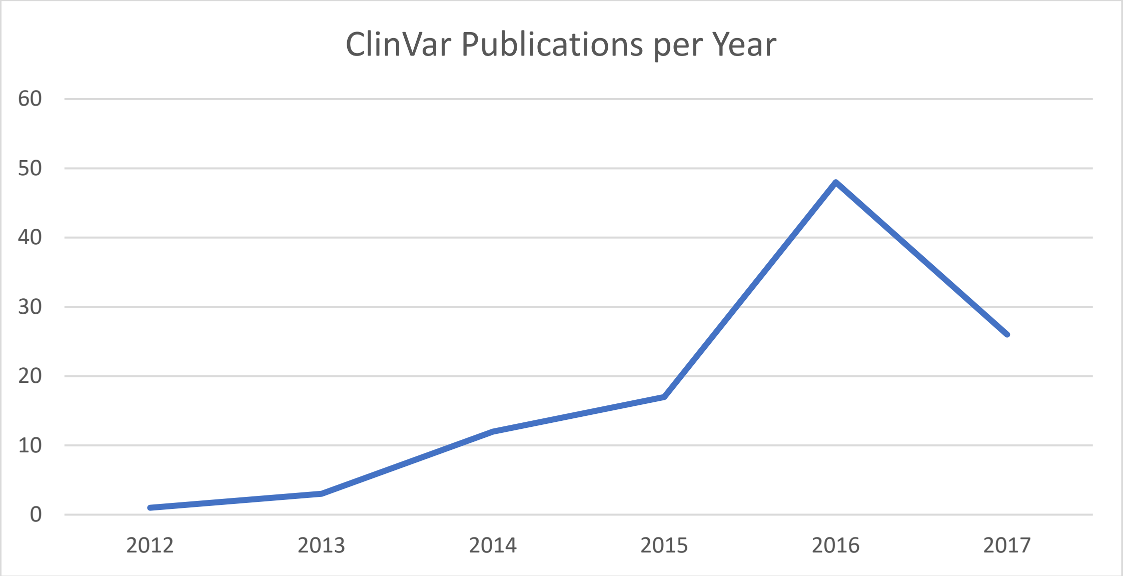


Twenty percent of the Visible Human Projects were reviewed in full-text to review how researchers who reference the Visible Human datasets are making a difference with their work. Of the 92 papers reviewed in full-text, 23 papers documented the creation of surgical simulations for improving training for surgeons and creating patient-specific simulations and models to create customized surgical plans, virtual reality applications to improve neurosurgical practices and knee arthroplasty training, and the creation of models that help detect and diagnose obstructive sleep apnea. These are only a few examples.

ClinVar

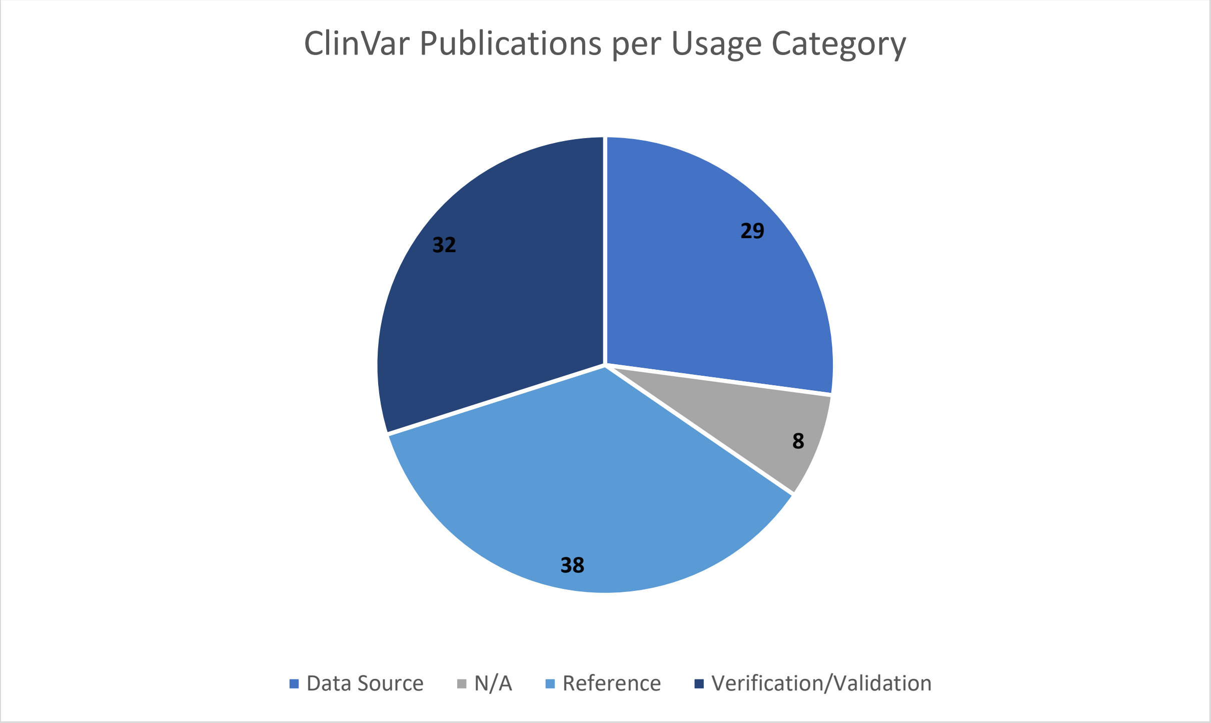
The ClinVar review encompassed 107 records after removing duplicates. ClinVar, released in beta in 2012 and in full in 2013, saw its highest number of references in 2016 (n=48, or 45%), possibly to be surpassed as the 2017 yield (n=26, or 24%) captured only the first third of the year. Based on the number of publications for the first third of 2017, an estimated projection of publications referencing ClinVar could be as much as 78 publications for the year 2017 which would surpass the 2016 total. An NLM author was the lead author on 11 of the 107 publications, or 10% of the publications.

Figure 6. ClinVar Publications per Year (2012-2017)



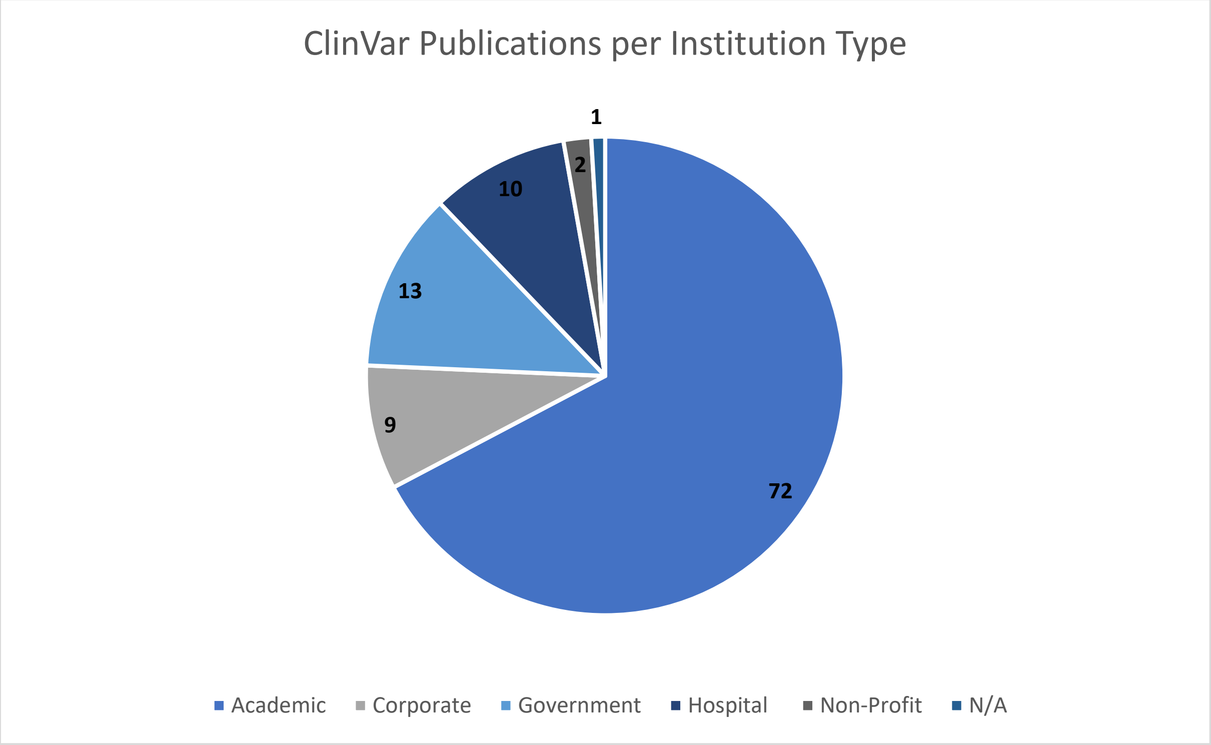
For ClinVar, the reported uses are evenly split between serving as an information reference resource (n=38, or 36%), a validation tool (n=32, or 30%), and as a data source for populating another resource (n=29, or 27%).

Figure 7. ClinVar Publications per Usage Category



The heaviest reported use of ClinVar is from academic institutions (n=72, or 67%), with a dramatic drop for government (n=13, or 12%), hospital (n=10, or 9%), and corporate users (n=9, or 8%).

Figure 8. ClinVar Publications per Institution Type



ClinVar has reported use from authors in 19 countries; outside of the US (n=64, or 60%) the country with the highest publication volume is China (n=10, or 9%). Within the US, publications came from 20 states with Maryland (n=16, or 25% of all US publications and 15% of all publications) and California (n=11, or 17% of all US publications or 10% of all publications) authoring the most papers.

Figure 9. ClinVar Publications per Country

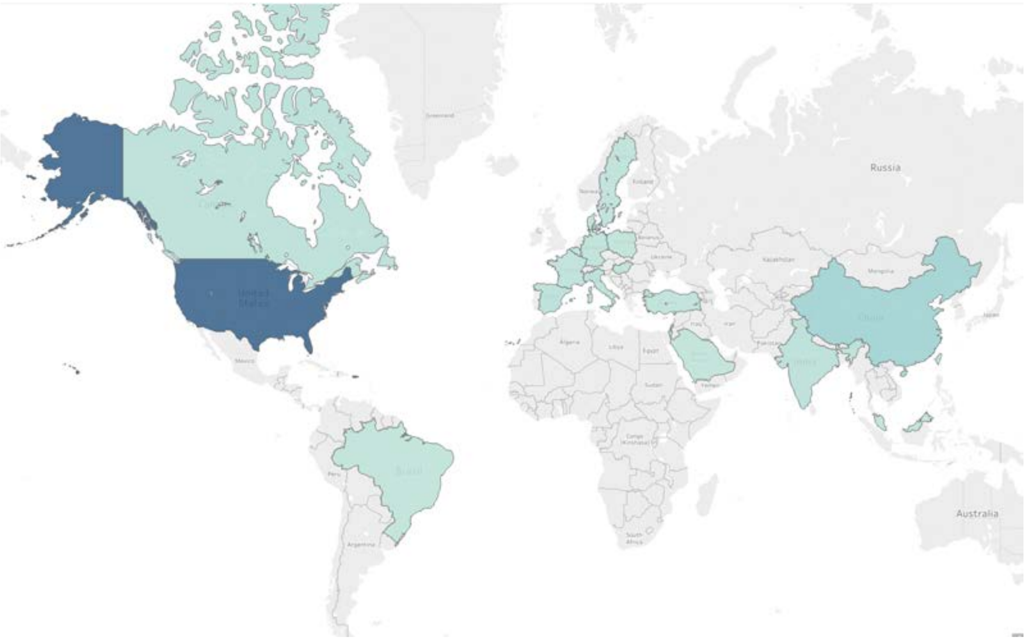
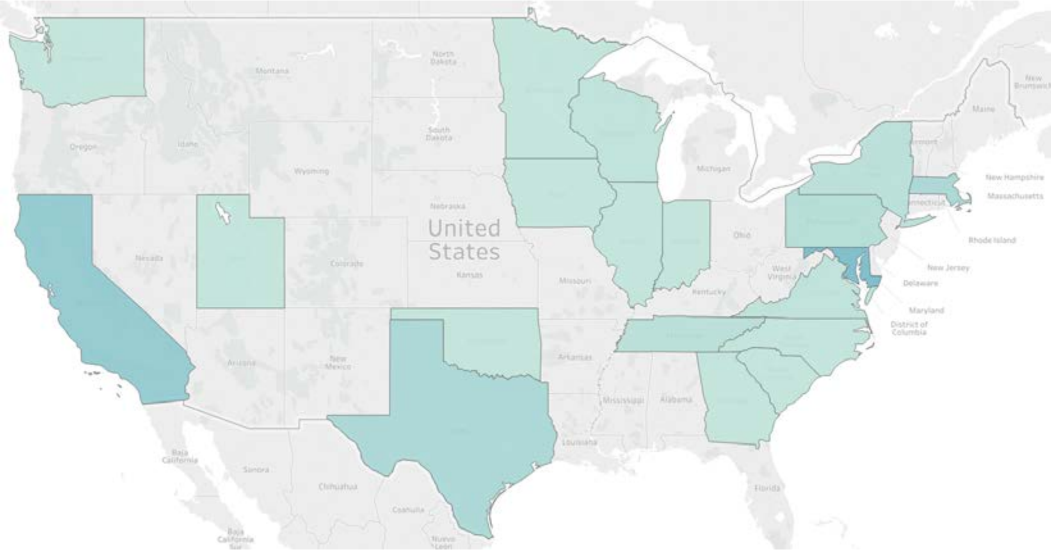


Figure 10. ClinVar Publications per US State



Twenty percent of the ClinVar results were reviewed in full-text to capture any indications of the impact ClinVar has had outside of NLM. Twenty papers were reviewed in full-text, of which ten had identifiable applications, software, and tools that have been made using ClinVar. Examples include Variation Viewer, Targeted Enrichment Analysis and Management (TEAM), an unnamed pharmacogenomic linked data project, the Scripps Genome ADVISER CNV, CryptSplice, BALL-SNP, GEMINI (GEnome MINIng), and MARRVEL (model organism aggregated resources for rare variant exploration).

Open-i and OSIRIS

Open-i and OSIRIS were included in the initial data extraction screening, but due to the challenges of ambiguous naming and how the bibliographic databases interpreted the search queries, most of the search results were marked non-relevant.

For Open-i, the original search yield after removing duplicates was 429. After first pass screening, 422 references were designated non-relevant leaving only five relevant references for this product. The five papers were all published in a two-year period, 2014-2016, and all within the US from authors located in either Maryland (n=4) or New York (n=1). The publications were mostly produced by authors affiliated with a government institution (n=4) or an academic institution (n=1). All the papers were review articles that described what Open-i does and providing recommendations on how to best use it.

OSIRIS had a substantial initial yield even after removing duplicate references. The 2,141 references were reviewed for relevance and reduced the list to a single relevant reference. This paper is a 2011 NCBI publication describing the new software and its release to the public. A key challenge with OSIRIS is that many other products use that same name. In addition to the open source software created by NLM, OSIRIS is also the name of a prominent banking software, an Egyptian god, a planet and an asteroid, the name of a serial publication, and is a serverless portal system for peer-to-peer file sharing.

Discussion

Bibliometric analysis can provide data that can be used as part of an evaluation of product impact: It can answer how many publications have been released citing a product, who is publishing those papers, where they are located, and to some extent how the products are being used and in naming what new tools, methods, or tests have been developed. The limitations of this method of analysis include basing the research exclusively in the published biomedical literature; social media, grey literature, and other resources were not included in the scope of this work. Bibliometric analysis can be used in conjunction with other usage metrics to evaluate the impact of a product.

Successful completion of this project required accurate product name recognition in the searches, and careful attention to the search query interpretations for each database. In the case of OSIRIS, the name ambiguity posed a significant challenge in identifying results to screen as the searches located thousands of unique results that were about non-NLM products. For Open-i, the databases incorrectly interpreted the search query despite the use of quotation marks to demark a specific phrase to search. These issues should be addressed in any future extensions of this work.

Many products and tools are built using NLM resources by academic research teams and corporate entities; this report highlights some examples. A future consideration is to address how any changes made to the NLM product will impact these secondary products. The secondary products and tools identified through this or future bibliometric investigations can serve as an entry point for establishing a relationship with these product developers. Opportunities for targeted usability and needs assessments could be established.

Possible future directions for continuing this work include a 1:1 comparison of traditional access usage metrics and web analytics data with the publication details available through bibliometric inquiry. While this project did not do extensive citation or network analysis, that is another avenue of continuation. Further explorations of the secondary products developed by using NLM resources is recommended next step, as is reviewing samples of the highest and lowest yield results in the initial PubMed scoping search for themes influencing the inclusion or lack of inclusion of NLM product names in the bibliographic records.

Acknowledgements

This project would not have been possible without my sponsor, Dina Demner Fushman. Her support and guidance throughout this process has been invaluable, from the early stages of developing the initial question and framing the methodology, through to the report writing stages. The work of the 2017 MeSH Indexing Impact Group led to the creation of this project; their inquiry into the uses and value of MeSH to various populations prompted the questions that led to this investigation of the use and impact of other NLM resources. Former Associate Fellow Kate Masterton's work in investigating uses of MeSH, as well as her encouragement and availability for questions, has been much appreciated.

Kathel Dunn and the 2016-2017 Associate Fellowship cohort (Megan Fratta, Kendra Godwin, and Tyler Moses) were instrumental in guiding my thinking for this project, encouraging me to not get lost in the data and better communicate the importance of my findings. Final thanks to the NLM leadership and Library Operations for making the Associate Fellowship program possible.

References

Beyond impact: measuring research, making a difference. (2017). Retrieved August 21, 2017, from http://beyond-impact.org/

Broadus, R. N. (1987). Early approaches to bibliometrics. *Journal of the American Society for Information Science (1986-1998)*, *38*(2), 127.

Defining impact. (2014). Retrieved August 21, 2017, from http://nilppa.newknowledge.org/the-white-paper/defining-impact/

Evaluation resources. (2017). Retrieved August 21, 2017, from https://www.imls.gov/research-evaluation/evaluation-resources

Hood, W. W., & Wilson, C. S. (2001). The Literature of bibliometrics, bcientometrics, and informetrics. *Scientometrics, 52*(2), 291. doi:10.1023/a:1017919924342

Small, H.G. (1977). Co-citation model of a scientific specialty: – a longitudinal study of collagen research. *Social Studies of Science*, 7 (2), 139–166.

Thomson Reuters. (2008). Using bibliometrics: a guide to evaluating research performance with citation data [White paper]. Retrieved August 21, 2017, from http://ip-science.thomsonreuters.com/m/pdfs/325133_thomson.pdf

Appendix 1: Duplicate Removal Counts

ClinVar

Database	Results	Results after removing duplicates
PubMed	90	90
IEEE Xplore	1	1
Scopus	83	7
Web of Science	77	9
LISTA	0	0
TOTAL	251	**107**

Open-i

Database	Results	Results after removing duplicates
PubMed	59	59
IEEE Xplore	17	14
Scopus	315	250
Web of Science	249	105
LISTA	1	1
TOTAL	641	**429**

OSIRIS

Database	Results	Results after removing duplicates
PubMed	230	227
IEEE Xplore	82	78
Scopus	1796	1463
Web of Science	1520	327
LISTA	55	46
TOTAL	3683	**2141**

TOXNET

Database	Results	Results after removing duplicates
PubMed	106	104
IEEE Xplore	0	0
Scopus	133	33
Web of Science	90	10
LISTA	39	34
TOTAL	368	**181**

Visible Human Project

Database	Results	Results after deduplication
PubMed	154	150
IEEE Xplore	66	48
Scopus	471	275
Web of Science	271	49
LISTA	17	11
TOTAL	979	**533**

Appendix 2: Random Sampling for Full-Text Review

For the full-text review to identify product impacts, 20% of relevant results were identified and reviewed; annotated copies of selected files are available in PDF form in Supplemental File D. The 20% of papers were identified by random sampling: Each reference was assigned a unique reference ID, then a random number generator was used to isolate the appropriate number of reference IDs. Open-i and OSIRIS had too few relevant results to undergo this evaluation.

ClinVar

Category	Article Counts
Total Results Reviewed	107
Relevant Results	99
20% of Relevant Results for Full Text Sampling	20
Impacts Identified	8

TOXNET

Category	Article Counts
Total Results Reviewed	181
Relevant Results	181
20% of Relevant Results for Full Text Sampling	36
Impacts Identified	8

Visible Human Project

Category	Article Counts
Total Results Reviewed	533
Relevant Results	460
20% of Relevant Results for Full Text Sampling	92
Impacts Identified	23