# Comparing Literature and the Real-World:

## using data science to track temporal trends

Shannon Sheridan

National Library of Medicine Associate Fellow

Date of Submission: August 22, 2018

Project Sponsors: Vojtech Huser

# Table of Contents

# Abstract

Objective: Investigate if public open data can be used to gain insights into temporal trends in medical literature and real-world data, demonstrating proof of concept. The goal was not to analyze the trends ourselves, but rather see if creating a semantic match was possible, and if that produced any output that could be analyzed.

Methodology: The project was split into two parts. The first half focused on work Dr. Vojtech Huser, a Lister Hill staff scientist and sponsor of the project, was doing with prescription drug data while also allowing the Associate the opportunity to learn the R programming language and practice with an expert. The second half of the project involved the Associate working with a different open access data set of performed medical procedures. The Associate conducted her own downloading and cleaning of the data, which involved removing improperly formatted or incomplete data and ensuring that column names reflected the content of the column. She then parsed the data and performed an analysis of that data. This work culminated in the creation of algorithms and proof of concept graphs that show open data can be used to investigate temporal trends. The data were presented in a poster presentation at the Mobilizing Computable Biomedical Knowledge Conference held at the National Library of Medicine on July 10-11.

Results: Concrete deliverables from the project included: a brief literature review, cleaned and graphed data sets for MeSH and medical procedures, a GitHub repository for the project, and a poster presented at a professional conference.

Conclusion: The project is the first study comparing literature and real-world datasets (for drugs or medical procedures). The Associate's data is public, providing researchers the opportunity to use it for analysis. Huser intends to use the algorithms and data sets thus far created for the prescription data and continue his investigation in that arena.

# Introduction

The practice of medicine is constantly changing as new drugs, therapeutic procedures, or diagnostic tests are introduced. Patients and family members unfamiliar with the latest research developments for a disease of interest may benefit if they can easily review trends in the medical literature and in medical practice. These trends can also be of interest to researchers. The Associate Fellow and Dr. Vojtech Huser, a staff scientist from Lister Hill, took advantage of open data initiatives that allow public access to data about medical literature (PubMed), delivered healthcare services, and drug prescriptions (Medicare data from Center for Medicaid and Medicare Services; aggregated data (not individual patient level)). The goal was to demonstrate that public open data can be used to gain insights into temporal trends in medical literature and real-world data. Throughout the course of the project, three data sets were used: the PubMed data set of MeSH terms, the drug prescription data set from Medicaid, and the delivered healthcare services data set from CMS.

This project was split in two parts. As the Associate had no prior knowledge of R, the programming language Huser had been using for the project, the first step was for the Associate to learn the basics of R programming. This practice work was done primarily on the drug prescription data set, Huser's focus of study. When the associate grew comfortable with her programming skills, she performed a search for other open data sets to compare to the original PubMed set. This search uncovered the medical procedures data set from the Centers for Medicare & Medicaid Services.

The Associate performed her own cleaning and analysis on this data set, and then distilled that information to create proof of concept graphs that show it is possible to create visual representation of the graphs for comparison. An example of those graphs was then used for a poster submission to the Mobilizing Computable Biomedical Knowledge Conference held at the National Library of Medicine on July 10-11.

Following best practices for research data management, the Associate created a GitHub repository for the purposes of data management and scientific openness. This repository, called DataTrends, only holds data and code related to the poster presentation and the Associate's work on the medical procedure data. Drug-related analyses are not included in this GitHub repository since there is ongoing analysis and improvements being made by Huser.

# Methodology

## Part 1-Prescription Drugs and Training:

The Associate and Huser held weekly meetings. The focus of these meetings was for the Associate to bring questions about the programming she was attempting to do, and then for the pair to discuss goals for each of them to work on in preparation for the next meeting. Much of the more difficult coding during this time was eventually done by Huser, to show the Associate how certain tasks could be accomplished in R.

The first step of the process was to download, parse, and clean the MeSH and prescription drug data sets. All of the processing work for the project was done with the R programming language, executed in RStudio, an integrated development environment for R. This was a multi-day process that Huser was already in the midst of completing when the Associate joined the team. The MeSH file was downloaded from the National Library of Medicine's website (https://www.nlm.nih.gov/mesh/filelist.html). The prescription drug data was downloaded from data.medicaid.gov (https://data.medicaid.gov/browse?category=State+Drug+Utilization&limitTo=datasets).

It was during this time the Associate completed the Johns Hopkins Data Science Specialization course on Coursera (https://www.coursera.org/specializations/jhu-data-science). This course was helpful not only in reinforcing basic R skills, but also introducing the Associate to some of the best practices in the realm of data science, such as commenting out code, maintaining version control, and standard file naming practice.

Huser also asked for the Associate to do a literature review and write "a two to three sentence summary" of the results (Appendix 1). This review was conducted in PubMed, keyword searching for articles about MeSH, selecting articles that discussed using MeSH in a non-indexing manner, and screening for those that discussed comparing MeSH terms and real-world data with semantic matching. The results of the summary conclude that while MeSH is being used for a variety of non-indexing purposes, there is no literature that compares available, real-world data for medical procedures or prescription drug usage to the literature domain in the context of a semantic match.

Using exact string matching, a semantic map was created between the MeSH data set and the prescription data, revealing 808 semantic matches, or list of identical entries, between the two data sets. This means that there were 808 occasions where a MeSH term had a corresponding occurrence in the drug prescription data. Code was written to create graph representations of those 808 drugs for a visual representation of trends over time. The project sponsor asked for an informal survey of the graphs and possible drugs for in-depth investigation. The resulting graphs of those 808 drugs in MeSH were manually examined by the Associate to investigate which showed the possibility of having interesting trends in the literature (Appendix 2). There were no quantitative requirements (e.g., increase by a certain percent over a specific period) for a drug to be included in this list.

Finally, the Associate investigated sources of other open access data sets that could possibly be compared to the MeSH data set, as the prescription drug data set had been. This search investigated both government data sets and those that are held by private institutions that are open to the public for download. One interesting data set, which spawned the Associates own mini-project, was found in the Center for Medicaid & Medicare Services data repository as “Medicare Provider Utilization and Payment Data: Physician and Other Supplier” (https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Physician-and-Other-Supplier.html).

## Part 2-CMS Medical Procedures:

The Associate created a GitHub repository for the purposes of data management and scientific openness for her mini-project, available at https://github.com/sheshan93/DataTrends. That project repository contains codes, result files, and graphs of the work the Associate conducted with the medical procedures data set.

The Associate downloaded, parsed, and cleaned the medical procedures data set. The CMS Medicare procedure data (available as Medicare National HCPCS Aggregate Summary Tables) consisted of 50,614 procedures spanning from 2012 to 2015. HCPCS (Healthcare Common Procedure Coding System) are used to document what procedures were performed and include Current Procedural Terminology (CPT) codes from the American Medical Association. For example, the HCPCS code for Rochephin, an injectable antibiotic, is J0696.

The Associate also used the PubMed data, which was already cleaned, from the first half of the project. The PubMed data consisted of 28,211 MeSH keywords (including MeSH Supplemental Concepts) and spanned from 1902 to 2017. The final cleaned version of the MeSH data is stored in the GitHub repository as “pubmed-trends-A.csv.”

The Associate took the dataset of 50,614 medical procedures and constructed an algorithm that identified only those procedures that demonstrated a 100% increase in use during the investigated period, creating a subset of 1,160 procedures. This was done in the hopes of discovering procedures with interesting upward trends that would be worth investigating further. That code is in the GitHub repository as “100-percent-procedures-increase-subset.R”

Using NLM’s Unified Medical Language System, we identified and mapped 2,827 procedures between the two data sets that are mapped to the same UMLS concept. This map is available in the GitHub repository as “mesh-cpt-map.csv”.

At the same time, the Associate continued working in R to create full graph sets of the medical procedures data and the MeSH term data (examples in Appendix 3). Both of the full sets of graphs are in the GitHub repository as “line-procedures-ratio-full.pdf” and “literature-stacked-graph-ratios-full.pdf”.

The Associate and Huser wrote and submitted a proposal for a poster presentation at the Mobilizing Computable Biomedical Knowledge Conference held at the National Library of Medicine on July 10-11 (Appendix 3). This proposal was based on the Associates work and analyses of trends with the medical procedures data set. The proposal was accepted, and the Associate created and displayed the poster for the conference (Appendix 4).

# Results

Concrete deliverables from the project include: a brief literature review, a poster presented at a professional conference, cleaned and graphed data sets for MeSH and medical procedures, and a GitHub repository for the project.  One important intangible benefit from the project is the Associate's gained knowledge of R programming.

As for research discoveries, the Associate found that it was possible to semantically map two different sets of data (MeSH and medical procedures) and retrieve meaningful data about trends from them. Huser and the Associate did learn that there are many shortcomings to the semantic matching process (contrasting our two separate projects), but it is possible to examine and compare two sets of data.

In order to provide this proof of concept, the Associate worked in R to create full graph sets of the medical procedures data and the MeSH term data, examples of which are below. These are examples of the medical procedure data graphs (Figure 1) and the literature data graphs (Figure 2).

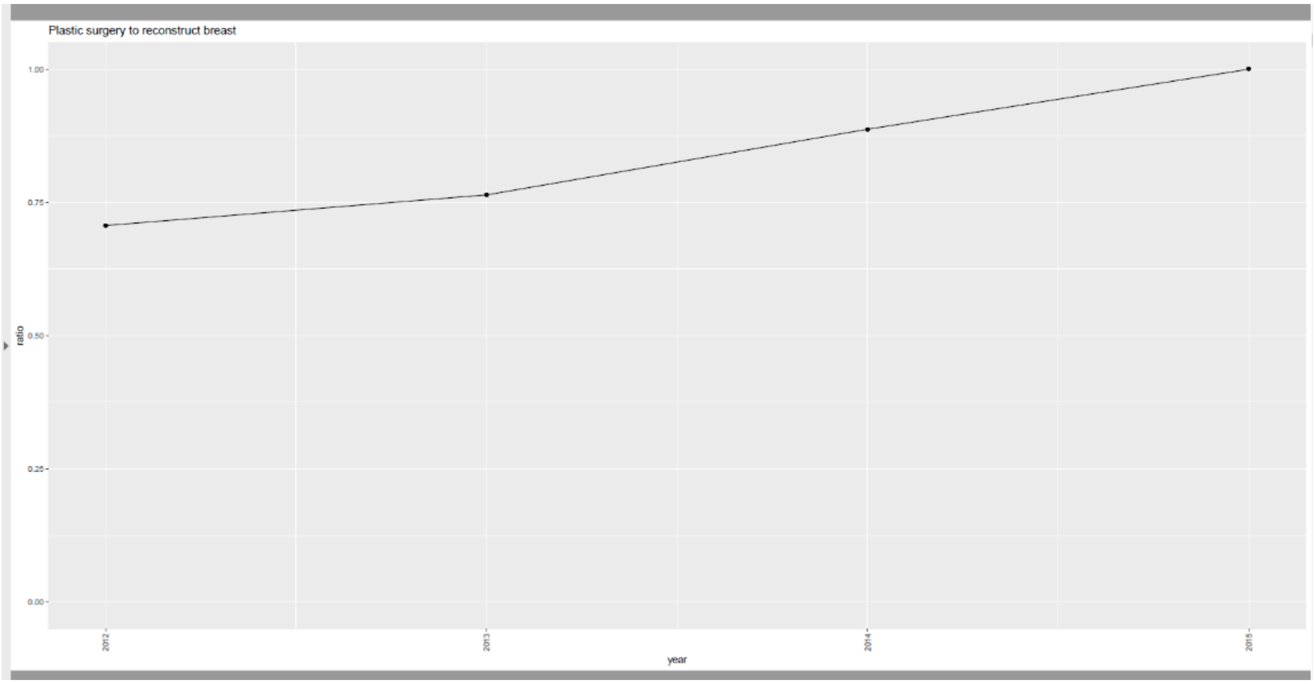


**Figure 1**

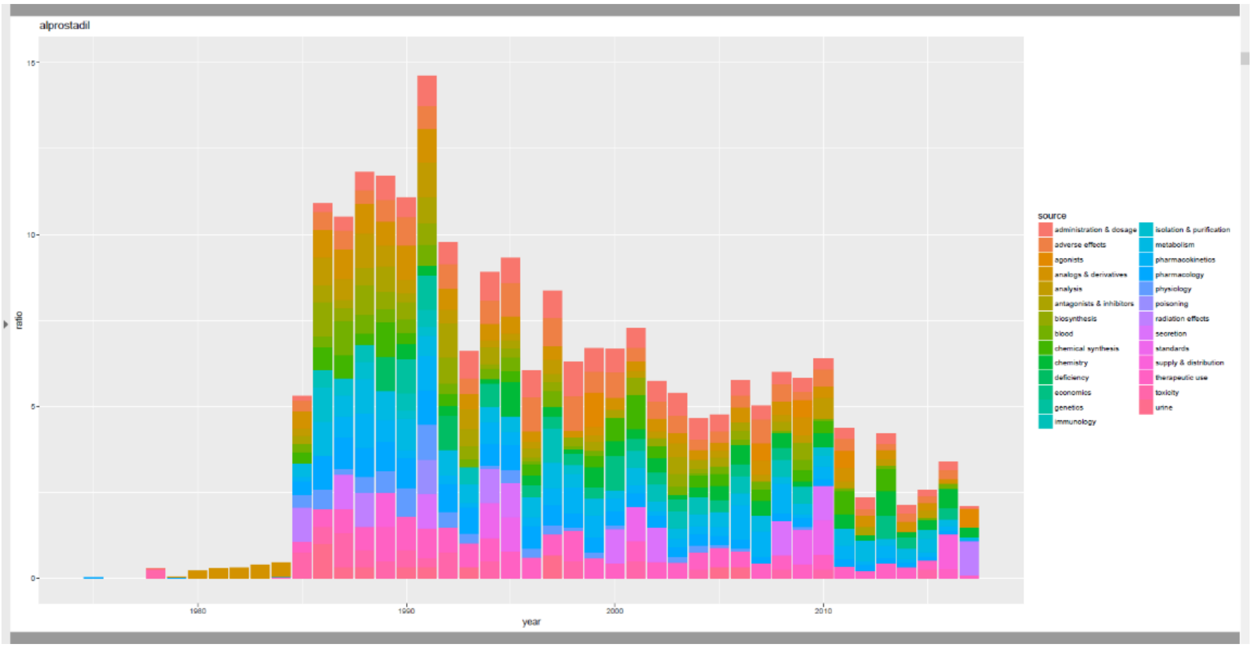


**Figure 2**

Both of the full sets of graphs are in the GitHub repository as "line-procedures-ratio-full.pdf" and "literature-stacked-graph-ratios-full.pdf". These graphs are the proof of concept, showing that open data sets can be mapped and represented visually to compare two data sets.

These trends can be examined by patients and practitioners alike as a tool that provides descriptive characterization of the usage of drugs or procedures, or by researchers interested in investigating certain drugs, procedures, or broad trends over time. There could be many practical benefits and academic insights from the having the final, graphed version of these data sets. The goal of the project was not to formally classify the trends, but rather see if creating a semantic match between two disparate data sets was possible, and if that produced any output that could be analyzed. The Associate and Huser showed that it is possible to take MeSH terms and match them with other data sets to allow for a comparison. For example, the poster presentation focused on the medical procedure "pericardiocentesis", and showed the graphs of literature usage, and the occurrence of it being billed by Medicare. Comparing the two graphs, one can see that there is a definite upward trend in the use of the term in MeSH indexing that is also present in the number of times the procedure was billed. However, some comparisons of other terms may show declining trends, contrasting trends between usage and literature, or no trend at all.

## Limitations:

CMS only provides medical procedural data since 2012, which limits the ability to detect long-term trends when comparing it to MeSH, which goes back as far as 1902. In addition, the Medicare data is comprised mostly of patients over 65 years old, is not normalized (for example, as a count of procedures per 10k patients), and is biased by changes in the total Medicare population over time. For medical literature, only articles with assigned MeSH terms were included in our study. Finally, within the medical procedure data, we observed differences in semantic granularity. That is, oftentimes there is not a 1:1 map that can be created between the medical procedure codes and corresponding index terms in MeSH.

# Future Work

As proof of concept has been achieved with the medical procedures data set, future work would be more analytical in nature, such as better methods for trend classification (strongly rising, declining, mixed, or no trend, as examples). Researchers could investigate the “why’s” of the trends, and their possible implications for biomedicine. Many things are possible, as the real-world procedure data is available to the public. This would allow a variety of audiences’ access to the results. For example, patients could investigate whether certain procedures they may be interested in are on an increase or a decrease, or if a certain disease they have is being researched more in literature. Practitioners could use the data for a similar purpose.

One possible area that interests the Associate would be to look at the data set and analyze the time lag between when a procedure is discussed in the medical literature, and when it begins to show a comparable trend in real-world usage. There is much discussion in the literature about the lag between publication and implementation in biomedicine. This data set could be another piece of the puzzle in examining how long such a gap is.

Huser plans to return to the prescription drug data set and continue his study of the temporal trends between that data and PubMed. Possible future work includes: replacing the exact string matching method with UMLS based mapping, creating criteria for identifying ingredients of interest for further analysis, and adding non-CMS drug usage data sources. Due to assisting the Associate and his other responsibilities, at this point not much more analysis has been done with the drug dataset.

# Conclusion

This comparison is the first such study that puts data from literature side by side with data from the real-world (for drugs or medical procedures). The project demonstrates that public open data can be used to gain insights into temporal trends in medical literature and real-world data. These trends can be examined by patients and practitioners alike to improve health outcomes, or for researchers investigating trends in literature and real-world usage. A map was created for both medical procedures and drug prescriptions that show semantic differences and similarities in terminology used by MEDLINE for literature and CMS for real-world data. Where there is a semantic match, one is able to compare trends in literature and the real-world. The merged literature and real-world data knowledge base allows further automated analyses, or manual analysis, into these trends.

# Appendix 1-Literature Review

In the literature, there is virtually no crossover between studies that examine trends in drug prescriptions and corresponding literature that contains those drugs as medical subject headings (MeSH).

Examining the trends of drug utilization, especially amongst particular subsets of a population, is common (Brady et al., 2014; de Hoyos-Alonso, Tapias-Merino, Meseguer Barros, Sánchez-Martínez, & Otero, 2015; François, Elbaz, & Pelletier Fleury, 2017; Kantor, Rehm, Haas, Chan, & Giovannucci, 2015; Regev, Eisenberg, Tansky, & Hadad, 2011; Villanueva, López de Argumedo, & Elizondo, 2016).

There is also literature using MeSH to study trends within the literature domain. Some examples are detecting emerging medical informatics research trends (Lyu, Yao, Mao, Zhang, 2015), studying MeSH co-occurrences (Kastrin, Rindflesch, & Hristovski, 2016; Miñarro-Giménez, Kreuzthaler, & Bernhardt-Melischnig, 2015; Miñarro-Giménez, Martínez, & Fernández-Breis, 2016;), or mining via data or text (Shan, Lu, Min, Que, & Zhang, 2016; Yea, Seong, Jang, & Kim, 2016; Xiang, Qin, Qin, & He, 2013).

However, in spite of the literature in these particular areas of study, there is no literature that compares available data to the literature domain in the context of a semantic mismatch. Our survey looks to relate literature usage of MeSH drug terms to actual usage of those drugs.

**Works Cited**

Brady, J. E., Wunsch, H., DiMaggio, C., Lang, B. H., Giglio, J., & Li, G. (2014). Prescription Drug Monitoring and Dispensing of Prescription Opioids. Public Health Reports, 129(2), 139–147.

de Hoyos-Alonso M.C., Tapias-Merino E., Meseguer Barros C.M., Sánchez-Martínez M., Otero A. (2015). Consumption trends for specific drugs used to treat dementia in the region of Madrid (Spain) from 2002 to 2012. Neurología 30(7), 416-24. 10.1016/j.nrl.2014.02.007

François M., Sicsic J., Elbaz A., Pelletier Fleury N. (2017) Trends in Drug Prescription Rates for Dementia: An Observational Population-Based Study in France, 2006-2014. Drugs & Aging, 34(9), 711-21. 10.1007/s40266-017-0481-7

Kantor, E. D., Rehm, C. D., Haas, J. S., Chan, A. T., & Giovannucci, E. L. (2015). Trends in Prescription Drug Use among Adults in the United States from 1999–2012. JAMA, 314(17), 1818–1831. http://doi.org/10.1001/jama.2015.13766

Kastrin A., Rindflesch T.C., Hristovski D. (2016). Link Prediction on a Network of Co-occurring MeSH Terms: Towards Literature-based Discovery. Methods of Information in Medicine, 55(4), 340-6. 10.3414/ME15-01-0108

Lyu P.H., Yao Q., Mao J., Zhang S.J. (2015), Emerging medical informatics research trends detection based on MeSH terms. Informatics for Health and Social Care, 40(3). https://doi.org/10.3109/17538157.2014.892490

Miñarro-Giménez J.A., Kreuzthaler M., Bernhardt-Melischnig J., Martínez-Costa C., Schulz S. (2015). Acquiring Plausible Predications from MEDLINE by Clustering MeSH Annotations. Studies in Health Technology and Informatics, 216, 716-720. 10.3233/978-1-61499-564-7-716

Miñarro-Giménez J.A., Martínez M.Q., Fernández-Breis JT.., Schulz S. (2016). Publishing Biomedical Predication Repository About MeSH Co-Occurrences in MEDLINE. Studies in Health Technology and Informatics, 228, 765-9. 10.3233/978-1-61499-678-1-765

Regev D., Eisenberg E., Tansky A., Hadad S. (2011). Opioid consumption in a tertiary hospital setting over an 8-year timeframe--a potential resource for tracking trends in pain management. Journal of Pain & Palliative Care Pharmacotherapy, 24(2), 113-20. 10.3109/15360288.2011.558992

Shan, G., Lu, Y., Min, B., Qu, W., & Zhang, C. (2016). A MeSH-based text mining method for identifying novel prebiotics. Medicine, 95(49), e5585. http://doi.org/10.1097/MD.0000000000005585

Villanueva G., López de Argumedo M., Elizondo I. (2016). Dementia drug consumption in the Basque Country between 2006 and 2011. Neurología 31(9), 613-19. 10.1016/j.nrl.2014.09.006

Yea S.J., Seong B., Jang Y., Kim C. (2016). A data mining approach to selecting herbs with similar efficacy: Targeted selection methods based on medical subject headings (MeSH). Journal of Ethnopharmacology, 182, 27-34. https://doi.org/10.1016/j.jep.2016.02.007

Xiang, Z., Qin, T., Qin, Z. S., & He, Y. (2013). A genome-wide MeSH-based literature mining system predicts implicit gene-to-gene relationships and networks. BMC Systems Biology, 7(Suppl 3), S9. http://doi.org/10.1186/1752-0509-7-S3-S9

# Appendix 2-List of drugs for potential in depth study

lubiprostone

bevacizumab

everolimus

pemetrexed

nebivolol

certolizumab pegol

emtricitabine

rilpivirine

cetuximab

adalimumab

insulin glargine

bortezomib

omalizumab

ranibizumab

acetohexamide

amphotericin b

ascorbic acid

atenolol

bromocriptine

captopril

chlorhexidine

chlorpromazine

cimetidine

cisplatin

demeclocycline

dobutamine

ethambutol

glucosamine

lincomycin

methyltestosterone

phenobarbital

phenytoin

epoprostenol

spironolactone

testosterone

thiamine

ticlopidine

ganciclovir

betaxolol

zalcitabine

isradipine

nedocromil

budesonide

insulin lispro (inverse)

# Appendix 3-Abstract Submission for Mobilizing Computable Biomedical Knowledge Conference

### Constructing a computable biomedical knowledge base for tracking the temporal trends of medical procedures in literature and real-world usage

Shannon Sheridan, Vojtech Huser; NIH NLM Bethesda, MD

**Introduction:** The practice of medicine is constantly changing as new drugs, therapeutic procedures, or diagnostic tests are introduced. Patients and family members unfamiliar with the latest research developments for a disease of interest may benefit if they can easily review trends in the medical literature and in medical practice. We took advantage of open data initiatives that allow public access to data about medical literature (PubMed database) and delivered healthcare services (Medicare data from Center for Medicaid and Medicare Services; CMS). For a learning health system, a computable knowledge base that relates research literature to real word data can be of value to machine reasoning.

**Methods:** For data about medical procedures, we extracted annual publication trends from PubMed and real-world usage trends from data.cms.gov. Counts of articles per MeSH (Medical Subject Heading) index term and counts of performed procedures were aggregated by calendar year. First, we analyzed literature trends and real-world data trends in isolation using their respective terminologies: MeSH for PubMed and Healthcare Common Procedure Coding System (HCPCS) for billed medical procedures. A large part of HCPCS terminology consists of a secondary, related procedural terminology called Current Procedural Terminology (CPT). Second, we compared the trends in literature and real-world data for medical procedures where the data was recorded at matching granularity—in other words, where there was a semantic match between MeSH and CPT. Our project repository at github.com/sheshan93/DataTrends contains additional methods, result files, and graphs (some of which are referenced below).

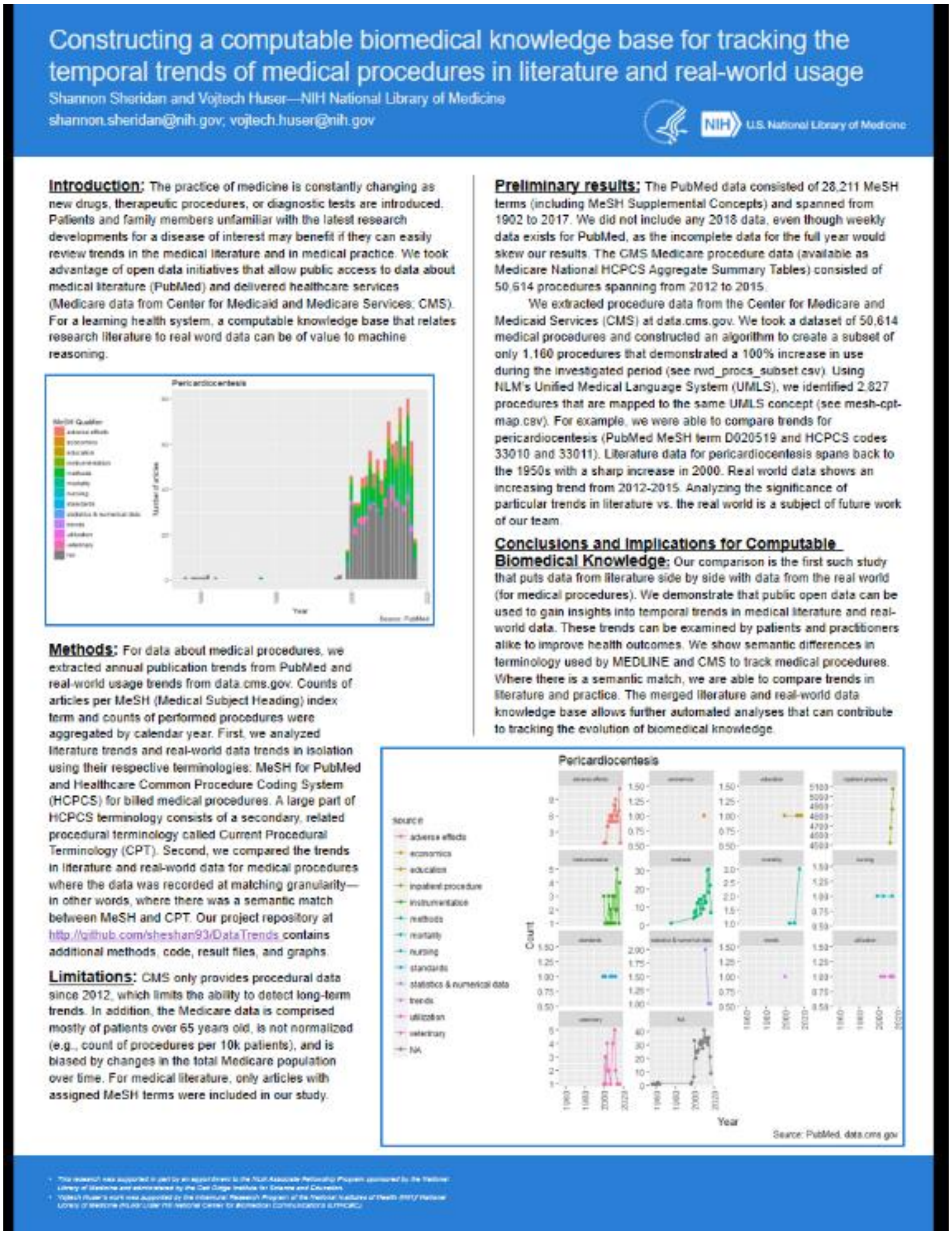
**Preliminary results:** The PubMed data consisted of 28,211 MeSH keywords (including MeSH Supplemental Concepts) and spanned from 1902 to 2017. We did not include any 2018 data, even though weekly data exists for PubMed, as the incomplete data for the full year would skew our results. The CMS Medicare procedure data (available as Medicare National HCPCS Aggregate Summary Tables) consisted of 50,614 procedures spanning from 2012 to 2015.

We extracted procedure data from the Center for Medicare and Medicaid Services (CMS) at data.cms.gov. We took a dataset of 50,614 medical procedures and constructed an algorithm to create a subset of only 1,160 procedures that demonstrated a 100% increase in use during the investigated period (see rwd_procs_subset.csv). Using NLM's Unified Medical Language System, we identified 2,827 procedures that are mapped to the same UMLS concept (see mesh-cpt-map.csv). For example, we were able to compare trends for pericardiocentesis (PubMed MeSH term D020519 and HCPCS codes 33010 and 33011). Literature data for pericardiocentesis spans back to the 1950s with a sharp increase in 2000. Real-world data shows an increasing trend from 2012-2015. Analyzing the significance of particular trends in literature vs. the real world is a subject of future work of our team.

**Limitations:** CMS only provides procedural data since 2012, which limits the ability to detect long-term trends. In addition, the Medicare data is comprised mostly of patients over 65 years old, is not normalized (e.g., count of procedures per 10k patients), and is biased by changes in the total Medicare population over time. For medical literature, only articles with assigned MeSH keywords were included in our study.

**Conclusions and Implications for Computable Biomedical Knowledge:** Our comparison is the first such study that puts data from literature side by side with data from the real world (for medical procedures). We demonstrate that public open data can be used to gain insights into temporal trends in medical literature and real-world data. These trends can be examined by patients and practitioners alike to improve health outcomes. We show semantic differences in terminology used by MEDLINE and CMS to track medical procedures. Where there is a semantic match, we are able to compare trends in literature and practice. The merged literature and real-world data knowledge base allows further automated analyses that can contribute to tracking the evolution of biomedical knowledge.

# Appendix 4-Poster for Mobilizing Computable Biomedical Knowledge Conference, Bethesda, MD, July 10-11



## Constructing a computable biomedical knowledge base for tracking the temporal trends of medical procedures in literature and real-world usage

Shannon Sheridan and Vojtech Huser—NIH National Library of Medicine
shannon.sheridan@nih.gov; vojtech.huser@nih.gov

NIH U.S. National Library of Medicine

**Introduction:** The practice of medicine is constantly changing as new drugs, therapeutic procedures, or diagnostic tests are introduced. Patients and family members unfamiliar with the latest research developments for a disease of interest may benefit if they can easily review trends in the medical literature and in medical practice. We took advantage of open data initiatives that allow public access to data about medical literature (PubMed) and delivered healthcare services (Medicare data from Center for Medicaid and Medicare Services; CMS). For a learning health system, a computable knowledge base that relates research literature to real word data can be of value to machine reasoning.

**Methods:** For data about medical procedures, we extracted annual publication trends from PubMed and real-world usage trends from data.cms.gov. Counts of articles per MeSH (Medical Subject Heading) index term and counts of performed procedures were aggregated by calendar year. First, we analyzed literature trends and real-world data trends in isolation using their respective terminologies: MeSH for PubMed and Healthcare Common Procedure Coding System (HCPCS) for billed medical procedures. A large part of HCPCS terminology consists of a secondary, related procedural terminology called Current Procedural Terminology (CPT). Second, we compared the trends in literature and real-world data for medical procedures where the data was recorded at matching granularity—in other words, where there was a semantic match between MeSH and CPT. Our project repository at http://github.com/sheshan93/DataTrends contains additional methods, code, result files, and graphs.

**Limitations:** CMS only provides procedural data since 2012, which limits the ability to detect long-term trends. In addition, the Medicare data is comprised mostly of patients over 65 years old, is not normalized (e.g., count of procedures per 10k patients), and is biased by changes in the total Medicare population over time. For medical literature, only articles with assigned MeSH terms were included in our study.

**Preliminary results:** The PubMed data consisted of 28,211 MeSH terms (including MeSH Supplemental Concepts) and spanned from 1902 to 2017. We did not include any 2018 data, even though weekly data exists for PubMed, as the incomplete data for the full year would skew our results. The CMS Medicare procedure data (available as Medicare National HCPCS Aggregate Summary Tables) consisted of 50,614 procedures spanning from 2012 to 2015.

We extracted procedure data from the Center for Medicare and Medicaid Services (CMS) at data.cms.gov. We took a dataset of 50,614 medical procedures and constructed an algorithm to create a subset of only 1,160 procedures that demonstrated a 100% increase in use during the investigated period (see rwd_procs_subset.csv). Using NLM's Unified Medical Language System (UMLS), we identified 2,827 procedures that are mapped to the same UMLS concept (see mesh-cpt-map.csv). For example, we were able to compare trends for pericardiocentesis (PubMed MeSH term D020519 and HCPCS codes 33010 and 33011). Literature data for pericardiocentesis spans back to the 1950s with a sharp increase in 2000. Real world data shows an increasing trend from 2012-2015. Analyzing the significance of particular trends in literature vs. the real world is a subject of future work of our team.

**Conclusions and Implications for Computable Biomedical Knowledge:** Our comparison is the first such study that puts data from literature side by side with data from the real world (for medical procedures). We demonstrate that public open data can be used to gain insights into temporal trends in medical literature and real-world data. These trends can be examined by patients and practitioners alike to improve health outcomes. We show semantic differences in terminology used by MEDLINE and CMS to track medical procedures. Where there is a semantic match, we are able to compare trends in literature and practice. The merged literature and real-world data knowledge base allows further automated analyses that can contribute to tracking the evolution of biomedical knowledge.

- This research was supported in part by an appointment to the NLM Associate Fellowship Program sponsored by the National Library of Medicine and administered by the Oak Ridge Institute for Science and Education.
- Vojtech Huser's work was supported by the Intramural Research Program of the National Institutes of Health (NIH)/ National Library of Medicine (NLM)/ Lister Hill National Center for Biomedical Communications (LHNCBC).