The OHDSI Community Dashboard: Tracking the Health and Impact of the Open Science Observational Health Data Sciences and Informatics Community

Star Liu, Asieh Golozar, Jody-Ann McLeggon, Adam Black, Paul Nagy

Background

Since 2014, the Observational Health Data Sciences and Informatics (OHDSI) community has continued to expand its network and body of literature in data standards[1], data characterization, safety surveillance, treatment effectiveness, risk prediction, and quality improvement. Together, the community accumulated over 300 publications on PubMed, 300 hours of YouTube content, 2,000 completed online courses (37% completion rate) on the European Health Data & Evidence Network (EHDEN) Academy, and hundreds of contributors have open sourced millions of lines of code on software development through GitHub. As the network continues to grow, there is a need for continuous evaluation of the entire ecosystem for aligning prospective research targets and efforts. We present here a web based graphical dashboard to help our community find articles, videos, and courses of interest and to track how we are making an impact in our field. We published the open-source framework (Community Dashboard: https://ohdsi.azurewebsites.net/) to study publication growth, identify key contributors, define clinical domains of interest, support network studies, and inform future OHDSI studies and collaborators' successes.

Methods

PubMed. The United States National Library of Medicine offers an open, free API service through its Entrez information retrieval system[2]. This service gives access to the PubMed database, where all articles were drawn. We validated a manual tracking effort based on keyword combinations. OHDSI articles were first identified through PubMed. Then, articles were matched using Levenshtein fuzzy matching algorithm, and corresponding citation counts were added using SerpAPI[3], which pulled citation counts from Google Scholar. Levenshtein matching also allowed us to identify new researchers entering OHDSI each year.

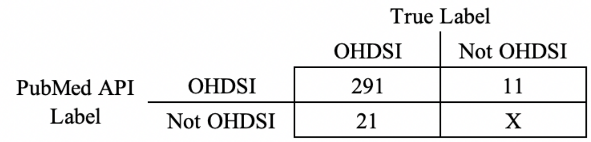
The goal was to accurately identify OHDSI articles with a high specificity. As such, our search strategy on PubMed included the search for terms "ohdsi," "omop," "observational health data sciences and informatics" (fixed phrase), "observational medical outcomes partnership" (fixed phrase), and "observational medical outcomes partnership common data model" (fixed phrase). Through PubMed, we used these search strategies to find articles with matching strings in the title, author, or abstract. We have created manual addition capabilities for edge cases.

YouTube and EHDEN. YouTube data API service gives access to metadata on each video's duration and viewership status. Although the API produces a different subset of relevant videos from the same search query, we developed a solution that only looks for new videos.

The EHDEN Academy, in collaboration with OHDSI, created online learning resources for researchers, developers, and analysts who are new to OHDSI. It hosts a data API service that gave access to metadata on the specialty of the courses and the prevalence of training completions.

Both PubMed and YouTube data API connections were coded in Python Version 3.7. The final scripts automatically run daily as Azure Functions on Azure Cloud to search for new articles, videos, courses, and the associated metadata. The final collection of articles and videos were stored as JSON objects in Azure CosmosDB. The final dashboard was built upon the Flask framework in the backend and HTML and Dash Apps in the frontend.

Results

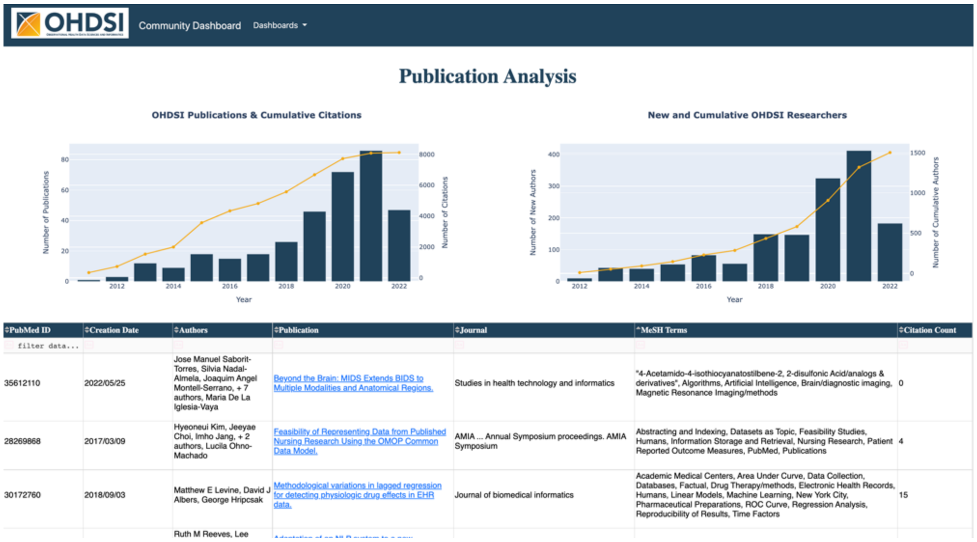


		True Label	
		OHDSI	Not OHDSI
PubMed API Label	OHDSI	291	11
	Not OHDSI	21	X

Figure 1. Confusion matrix: PubMed API label vs. true label. The number of non-OHDSI articles is undefined marked with an "X."

To validate our PubMed curation, we manually checked the results against a pre-validated list of OHDSI articles. As shown in Figure 1, the PubMed search strategy identified 302 articles on PubMed with 96.36% precision and

93.27% sensitivity. Among the 302 articles, 11 were not truly OHDSI-related and were flagged as false positives. The false positives were then manually removed from the collection of articles to track. Compared to the pre-validated collection of 217 articles, 21 of them were left out, leaving us with 21 “false negatives.” The false negatives were then manually added to the total collection of articles for tracking. This collection is stored on Azure CosmosDB and presented on the publication dashboard in a Dash data table.



Publication Analysis

OHDSI Publications & Cumulative Citations

New and Cumulative OHDSI Researchers

PubMed ID	Creation Date	Authors	Publication	Journal	MeSH Terms	Citation Count
35612110	2022/05/25	Jose Manuel Saborit-Torres, Silvia Nadal-Almela, Joaquim Angel Montell-Serrano, + 7 authors, Maria De La Iglesia-Vaya	Beyond the Brain: MIDS Extends BIDS to Multiple Modalities and Anatomical Regions.	Studies in health technology and informatics	"4-Acetamido-4-isothiocyanatostilbene-2, 2-disulfonic Acid/analogs & derivatives", Algorithms, Artificial Intelligence, Brain/diagnostic imaging, Magnetic Resonance Imaging/methods	0
28269868	2017/03/09	Hyeoneui Kim, Jeeyae Choi, Imho Jang, + 2 authors, Lucila Ohno-Machado	Feasibility of Representing Data from Published Nursing Research Using the OMOP Common Data Model.	AMIA ... Annual Symposium proceedings. AMIA Symposium	Abstracting and Indexing, Datasets as Topic, Feasibility Studies, Humans, Information Storage and Retrieval, Nursing Research, Patient Reported Outcome Measures, PubMed, Publications	4
30172760	2018/09/03	Matthew E Levine, David J Albers, George Hripcsak	Methodological variations in lagged regression for detecting physiologic drug effects in EHR data.	Journal of biomedical informatics	Academic Medical Centers, Area Under Curve, Data Collection, Databases, Factual, Drug Therapy/methods, Electronic Health Records, Humans, Linear Models, Machine Learning, New York City, Pharmaceutical Preparations, ROC Curve, Regression Analysis, Reproducibility of Results, Time Factors	15
		Ruth M Reeves, Lee	Adaptation of an NLP system to a new			

Figure 2. PubMed Publication dashboard highlights scholarship generated the using OMOP Common Data Model, OHDSI tools, or the OHDSI network.

As shown in the publication dashboard (Figure 2), incremental and cumulative citation and author counts are presented in line graphs overlaying bar charts. The curated collection of articles is presented in Dash data tables. Each article has citation counts, affiliated authors, and a set of Medical Subject Headings (MeSH) terms that can be mapped to SNOMED concepts. Researchers can easily find OHDSI publications to use for guidance and reference. Using citation count and authorship, we can identify key contributors and new emerging researchers/institutional partners. Mapping MeSH terms to SNOMED codes allows us to categorize each article based on the disease specification or the clinical domain. With this pipeline, we can find papers that share similarities, define areas of research that OHDSI has excelled, and help conduct observational research.

We used the same Dash app interface scheme for YouTube and EHDEN. YouTube content creation as well as EHDEN Academy course offerings and completion rates highlight the types of resources that are most sought after and valued by the community learners. Gauging the successes and failures of educational contents help create additional resources tailored to the interests of the members.

Conclusion

We built an open-source framework for harvesting data APIs from PubMed, YouTube, and EHDEN to track the health and development of OHDSI. Progress is underway to integrate GitHub analytics to measure the engagement of contributors. It would allow us to identify outstanding individuals and institutional contributions as well as those who need additional support from within the network. Together, this open-source tool would provide a holistic evaluation of the scholarship dissemination, educational support, and engagement of the network, serving as a measuring stick for success.

References

1. OHDSI – Observational Health Data Sciences and Informatics [Internet]. [cited 2022 Feb 15]. Available from: https://www.ohdsi.org/
2. Bio.Entrez package — Biopython 1.76 documentation [Internet]. [cited 2022 Feb 9]. Available from: https://biopython.org/docs/1.76/api/Bio.Entrez.html
3. Google Scholar API | Scrape Google Scholar - SerpApi [Internet]. [cited 2022 Feb 15]. Available from: https://serpapi.com/google-scholar-api