**Appendix**

# Project Repository

<https://lhncbc.github.io/CRI/ArticleLinkage>

# PubMed vs MEDLINE vs. PubMedCentral

PubMed is a database of biomedical literature citations produced and maintained by the National Library of Medicine (NLM). PubMed facilitates searching across multiple NLM literature resources, including PubMed Central (PMC), an open access archive of biomedical text, and MEDLINE, a curated database of biomedical literature citations that have been reviewed and recommended by a selection committee. See an extensive explanation of the differences between PubMed, MEDLINE, and PubMed Central at <https://www.nlm.nih.gov/bsd/difference.html>

MEDLINE articles are reviewed by NLM indexers and curated with funding, genetic, chemical and other metadata. In 2019, nearly 1.4 million citations were added to PubMed of which 1 million are included in MEDLINE.

# Secondary Source ID (SI) Field

The Secondary Source ID (SI) field contains information about types of data discussed in MEDLINE articles. Originally, only NLM indexers could modify the SI field meaning that only MEDLINE articles contained SI metadata. SI metadata was limited to databases included in the allowed list of [MEDLINE databank sources](https://www.nlm.nih.gov/bsd/medline_databank_source.html). However, publishers gained the ability to directly edit their own metadata in PubMed and began adding content to the SI field with the shift to the PubMed Data Management (PMDM) system in October 2016.1 NLM indexers created a system to identify the value manual citation review added for each journal in MEDLINE and identified journals most likely to include database content. The implementation of this processing system means that only approximately 45% of MEDLINE citations are manually reviewed by NLM indexers.

Example of an article that includes database (BioProject) but that database was not included in Associated Data (SI) field: <https://pubmed.ncbi.nlm.nih.gov/31034726/>

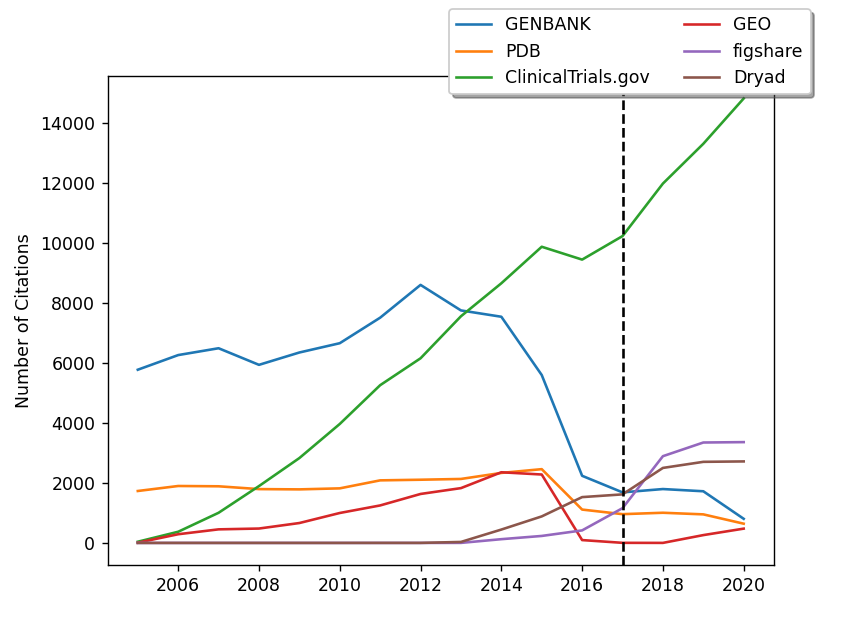
# XML Standards

PubMed only accepts citation metadata from publishers and other data providers that is formatted in the [PubMed XML Tagged Format](https://www.ncbi.nlm.nih.gov/books/NBK3828/#publisherhelp.PubMed_XML_Tagged_Format). Since the release of the PubMed Data Management System (PMDM) in October 2016, publishers and other data providers are responsible for accurately presenting their own citations in PubMed.

# SI Usage

## Visualization of Select Databases 2005 – 2020

GENBANK, Protein Data Bank (PDB), ClinicalTrials.gov, and Gene Expression Omnibus Database (GEO) are NIH/NLM databases tracked in the SI field. Dryad and figshare are databases identified as priorities for NLM indexers to review and add to article metadata in 2016.



## SI Citation Counts in PubMed by Year

This table shows SI citation counts in PubMed following the 2016 policy change that allowed publishers to edit the XML for the SI field and resulted in NLM indexers conducting less manual review of article metadata. For a complete view of the data from 2000-2020, go to <https://github.com/lhncbc/CRI/blob/master/ArticleLinkage/SI-stats-yearly.xlsx>.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Abbreviation** | **Name** | **2016** | **2017** | **2018** | **2019** | **2020** |
| ClinicalTrials.gov | ClinicalTrials.gov Database (NIH/NLM) | 9455 | 10240 | 11993 | 13320 | 14834 |
| ISRCTN | International Standard Randomized Controlled Trial Number (ISRCTN.org) | 807 | 584 | 670 | 675 | 494 |
| ANZCTR | Australian New Zealand Clinical Trials Registry | 459 | 349 | 413 | 436 | 343 |
| NTR | The Netherlands National Trial Register | 234 | 185 | 203 | 203 | 131 |
| EudraCT | EU Clinical Trials Register | 221 | 176 | 208 | 229 | 178 |
| ChiCTR | Chinese Clinical Trials Registry | 127 | 100 | 161 | 216 | 197 |
| DRKS | German Clinical Trials Register | 108 | 84 | 134 | 161 | 172 |
| UMIN CTR | University Hospital (Japan Registry Member) | 47 | 70 | 155 | 148 | 172 |
| IRCT | Iranian Registry of Clinical Trials | 55 | 39 | 73 | 90 | 78 |
| JPRN | Japan Primary Registries Network | 128 | 28 | 37 | 41 | 16 |
| CTRI | Clinical Trials Registry - India | 43 | 30 | 54 | 45 | 57 |
| PACTR | Pan African Clinical Trial Register | 31 | 28 | 40 | 24 | 28 |
| CRiS | Clinical Research Information Service, Republic of Korea | 22 | 19 | 29 | 30 | 41 |
| ReBec | Brazilian Clinical Trials Registry | 11 | 7 | 9 | 10 | 11 |
| TCTR | Thai Clinical Trials Registry | 5 | 2 | 10 | 14 | 10 |
| JapicCTI | Japan Pharmaceutical Information Center (Japan Registry Member) | 4 | 0 | 6 | 7 | 7 |
| SLCTR | Sri Lanka Clinical Trials Registry | 2 | 8 | 3 | 1 | 2 |
| RPCEC | Cuban Public Registry of Clinical Trials | 2 | 2 | 0 | 1 | 0 |
| JMACCT | Center for Clinical Trials, Japan Medical Association (Japan Registry Member) | 1 | 0 | 0 | 1 | 1 |
| REPEC | Peruvian Clinical Trial Registry | 0 | 0 | 0 | 0 | 0 |
| GENBANK | GenBank Nucleic Acid Sequence Database (NIH/NLM) | 2240 | 1687 | 1798 | 1724 | 805 |
| PDB | Protein Data Bank (NIH/NLM) | 1113 | 960 | 1007 | 953 | 642 |
| GEO | Gene Expression Omnibus Database (NIH/NLM) | 95 | 3 | 1 | 261 | 477 |
| Dryad | Dryad | 1529 | 1620 | 2503 | 2706 | 2720 |
| figshare | figshare | 418 | 1160 | 2895 | 3351 | 3364 |
| RefSeq | Reference Sequence Database (NIH/NLM) | 57 | 44 | 30 | 19 | 83 |
| OMIM | Mendelian Inheritance in Man (McKusick) | 1 | 0 | 0 | 0 | 0 |
| SRA | Sequence Read Archive (SRA) Database (NIH/NLM) | 36 | 0 | 0 | 31 | 38 |
| SWISSPROT | Protein Sequence Database | 8 | 19 | 28 | 8 | 0 |
| PubChem-Substance | PubChem Substance Database (NIH/NLM) | 131 | 147 | 36 | 0 | 0 |
| BioProject | BioProject Database (NIH/NLM) | 14 | 0 | 2 | 2 | 1 |
| PIR | Protein Identification Resource | 3 | 2 | 3 | 1 | 0 |
| GDB | Johns Hopkins University Genome Data Bank (ceased in 2008) | 0 | 0 | 0 | 0 | 0 |
| dbGaP | Database of Genotypes and Phenotypes (NIH/NLM) | 2 | 0 | 0 | 5 | 6 |
| dbSNP | Database of Short Genetic Variations (NIH/NLM) | 0 | 0 | 0 | 0 | 0 |
| UniProtKB | UniProt Knowledgebase Database | 0 | 0 | 0 | 0 | 0 |
| PubChem-BioAssay | PubChem BioAssay Database (NIH/NLM) | 1 | 0 | 0 | 0 | 0 |
| dbVar | Database of Genomic Structural Variation (NIH/NLM) | 0 | 0 | 0 | 0 | 0 |
| UniRef | UniProt Reference Clusters Database | 0 | 0 | 0 | 0 | 0 |
| PubChem-Compound | PubChem Compound Database (NIH/NLM) | 1 | 0 | 0 | 0 | 0 |
| UniMES | UniProt Metagenomic and Environmental Sequences Database | 0 | 0 | 0 | 0 | 0 |
| UniParc | UniProt Archive Database | 0 | 0 | 0 | 0 | 0 |

# References

Gollner K, Canese K. PubMed: Redesigning citation data management. In: *Journal Article Tag Suite Conference (JATS-Con) Proceedings 2017 [Internet].* Bethesda (MD): National Center for Biotechnology Information (US); 2017. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK425541/>

PubMed Help. Bethesda (MD): National Center for Biotechnology Information (US); 2005-. XML Help for PubMed Data Providers. 2008 Feb 26 [Updated 2020 May 12]. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK3828/>

<https://pubmed.ncbi.nlm.nih.gov/23874614/>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2663789/>