

PKG 2.0 Dataset Description

The tables starting with the beginning of A are the original data tables of PubMed. The tables with the beginning of B are the tables associated with external data sources to provide a support for PKG2.0. The table with the beginning of C is a table synthesized by extracting key information from the tables A and B for subsequent statistics and calculations.

The original document of each field description of PubMed can be found at:

https://www.nlm.nih.gov/bsd/licensee/elements_descriptions.html

1. A01_Articles

Specific information for each article.

Column Name	Description
id	Automatically incremented identifier.
PMID	Unique ID assigned by PubMed to identify PubMed articles.
PMID_Version	PMID version number, which is the record corresponding to all PMID versions in this table. That is, if the PMID version number is different, there are different records in this table, and the id field of each record is different. In other related article tables (such as C01_Article_simple), only the record with the highest version number is kept.
MedlineCitation_Owner	Organization responsible for creating and verifying citations, including: NLM, NASA, PIP, KIE, HSR, HMD, SIS, NOTNLM
MedlineCitation_Status	It is the stage of the article. There are seven possible values: Completed In-Process PubMed-not-MEDLINE In-Data-Review Publisher MEDLINE OLDMEDLINE
Journal_JournalIssue_PubDate_Year	Publication year of the current article.
Journal_JournalIssue_PubDate_MedlineDate	Publication year of the current article. If <i>Journal_JournalIssue_PubDate_Year</i> is null, the publication year of the current article is recorded in this field

2. A02_AuthorList

Specific information for each author.

Column Name	Description
PMID	Unique ID assigned by PubMed to identify PubMed articles.
Au_Order	Author order of the current authors in the author list of current articles.
LastName	Last name of the current author.
ForeName	Current author's name excluding the last name and suffix
LastNameForeName	MD5 code generated by the LastName and the first letter of ForeName
AuthorNum	Co-author number of the current article

Vetle_aid	Unique author ID allocated by Vetle.
S2ID	Unique author ID allocated by Semantic Scholar
AID	Unique author ID (The final author disambiguation result).

3. A03_KeywordList

Article keyword information: keyword information in this table is provided by the data producer.

Column Name	Description
PMID	Unique ID assigned by PubMed to identify PubMed articles.
Keyword	Keywords of the current article

4. A04_Abtract

The abstract of each article.

Column Name	Description
PMID	Unique ID assigned by PubMed to identify PubMed articles.
AbstractText	The abstract of the current Article

5. A05_GrantList

Grants details of each article.

Column Name	Description
PMID	Unique ID assigned by PubMed to identify PubMed articles.
GrantID	Funding identifiers, including research grant numbers or contract numbers (or both) that are financially supported by the US Public Health Service or any agency of the National Institutes of Health (NIH)
Project_Number	The NIH project number corresponding to the funded article. Each article may be funded by multiple projects. Multiple project numbers are separated by ",".

6. A06_MeshheadingList

Mesh Heading details of each article.

Mesh Heading refers to the NLM control vocabulary and medical subject heading (MeSH®), which is used to characterize the content of the articles represented by MEDLINE citations.

Column Name	Description
PMID	Unique ID assigned by PubMed to identify PubMed articles.
DescriptorName	Descriptors extracted from each article
DescriptorName_MajorTopicYN	If the MeSH descriptors assigned to the article is the Key word of the article, the value is Y, otherwise it is N.
DescriptorName_UI	Identify MeSH's unique encoding for each descriptor and qualifier
QualifierName	Qualifiers, including numbers and words

7. A07_SupplMeshList

Supplementary conceptual terms and protocol terms for each article.

Column Name	Description
PMID	Unique ID assigned by PubMed to identify PubMed articles.
SupplMeshName	Supplementary conceptual term
SupplMeshName_Type	The type of supplementary conceptual term
SupplMeshName_UI	MeSH unique identifier for supplementary protocols and diseases

8. A08_ChemicalList

The chemical substances and registry number covered in each article.

Registry Number refers to a code assigned by Chemical Abstracts Service to a specific chemical substance.

Column Name	Description
PMID	Unique ID assigned by PubMed to identify PubMed articles.
RegistryNumber	Unique Identifier to a specific chemical substance assigned by Chemical Abstracts Service
NameOfSubstance	The name of the specific chemical substance

9. A09_CommentsCorrectionsList

Reference information for each article, including the source, type, and PMID of the reference.

Column Name	Description
PMID	Unique ID assigned by PubMed to identify PubMed articles.
RefSource	Reference sources
RefType	Reference type
RefPMID	The PMID of the reference
RefNote	Corrections to records with incorrect references

10. A10_DatabankList

The search number of the molecular sequence database that appears in the PubMed article. The search number can find the information of the corresponding chemical molecule from the established molecular sequence database, avoiding the use of lengthy molecular formulas and graphics in the article.

Column Name	Description
PMID	Unique ID assigned by PubMed to identify PubMed articles.
DataBankName	Name of the molecular sequence database

11. A11_PersonalNameSubjectList

Individuals' names appear in <PersonalNameSubject> for citations that contain a biographical note or obituary, or are entirely about the life or work of an individual or individuals. Data is entered in the same format as A02_AuthorList.

12. A12_InvestigatorList

Each article corresponds to the NASA-funded principal investigator (PI) information, and they participated in the discussion and research of the article (but not necessarily the author).

Column Name	Description
PMID	Unique ID assigned by PubMed to identify PubMed articles.
AffiliationInfo_Affiliation	The affiliation which the researchers belong to
AffiliationInfo_Identifier	Unique identifier of the affiliation

13. A13_AffiliationList

Extracted affiliation information.

Column Name	Description
PMID	Unique ID assigned by PubMed to identify PubMed articles.
Au_Order	Author order of the current authors in the author list of current articles.
Affiliation	Affiliation string.
Affiliation_order	For authors belonging to multiple affiliations, use numbers to identify different affiliations (in no particular order)

14. A14_ReferenceList

Reference information (we use the information in this table to generate the author's self-cited record during the disambiguation of strong features).

Column Name	Description
PMID	Unique ID assigned by PubMed to identify PubMed articles.
RefRank	Reference ranking
RefCitation	Reference sources (journals, year of publication, etc.)
RefArticleId	The PMID of the reference
RefIdType	Reference source database (e.g., PubMed)

15. B01_Descriptor

Specific information of Descriptor of each article (used for the classification of articles), [Download MeSH Data](#) from NIH.

Column Name	Description
DescriptorUI	The unique identifier of the descriptor
DescriptorName	The unique term of the descriptor
DescriptorClass	The corresponding classification of the descriptor (such as subject index, document type)
DateCreated	The date of the recording
DateEstablished	The date available for retrieval
Annotation	Text information corresponding to the descriptors
HistoryNote	History of text messages designed to help online searchers
NLMClassificationNumber	The NLM classification numbers assigned to the index terms
OnlineNote	Text messages designed to help online searchers
PublicMeSHNote	Change history of information
ConsiderAlso	Cross-reference, point to similar descriptors
TreeNumberList	One or more sets of elements in the descriptor or qualifier record

16. B02_projects

Contains NIH funding information, Download link: <https://reporter.nih.gov/exporter>.

Column Name	Description
PMID	Unique ID assigned by PubMed to identify PubMed articles.
PubYear	Year the article was published.
AuthorNum	The number of authors of this article.
CitedCount	The number of times the article was cited (calculated from the C04_ReferenceList)

17. C01_Papers

Simple table of A01_Articles.

Column Name	Description
PMID	Unique ID assigned by PubMed to identify PubMed articles.
PubYear	Year the article was published.
AuthorNum	The number of authors of this article.
CitedCount	The number of times the article was cited (calculated from the C04_ReferenceList)

18. C02_Link_Papers_Authors

Simple table of A02_AuthorList.

Column Name	Description
PMID	Unique ID assigned by PubMed to identify PubMed articles.
PubYear	Year the article was published.
AuthorNum	The number of authors of this article.
Au_Order	Author order of the current authors in the author list of current articles.
AID	Unique author ID (The final author disambiguation result).
BeginYear	The year this author published the first paper.
CurrentAge	Author's academic age: CurrentAge=PubYear-BeginYear+1

19. C03_Affiliations

Merging Vetle_Map and A13_AffiliatioinList information, including the parsed organization information, such as Zip code, Location, Country, etc.

Column Name	Description
Id	Unique ID assigned to each record.
PMID	Unique ID assigned by PubMed to identify PubMed articles.
Au_Order	Author order of the current authors in the author list of current articles.
AffiliationOrder	Affiliation order of one author, in case one author has multiple affiliations.
Affiliation	Affiliation text.
Department	Department parsed from Affiliation.
Institution	Institution parsed from Affiliation.
Email	Email parsed from Affiliation.
Zipcode	Zipcode parsed from Affiliation.
Location	Location parsed from Affiliation.
Country	Country parsed from Affiliation.
AID	Unique author ID.
City	City parsed from Affiliation.
State	State parsed from Affiliation.
Vetle_Country	Country abbreviation.
Type	Affiliation type, such as COM, EDU, ORG, et.al.
Lat	Latitude of the Affiliation.
Lon	Longitude of the Affiliation.
Fips	Fips code.

20. C04_ReferenceList_Papers

The sources of data integration include PubMed's own citation data, NIH Open Citation Collection(run by iCite) and the OpenCitations Index of Crossref open DOI-to-DOI citations(run by OpenCitations).

Column Name	Description
Id	Unique ID assigned to each record.
PMID	Unique ID assigned by PubMed to identify PubMed articles.
RefArticleId	PMIDs of references.

21. C05_PIs

According to PMID and the author's name (full name of the last name and initials), match the author AID and NIH Principal Investigator (PI) number PIID to generate correspondence table C05, including PIID, AID, CORE_PROJECT_NUMBER, PIID, Application_ID, etc.

Column Name	Description
Id	Unique ID assigned to each record.
PI_ID	A unique identifier for each of the project Principal Investigators. Each PI in the RePORTER database has a unique identifier that is constant from project to project and year to year, but changes may be observed for investigators that have had multiple accounts in the past, particularly for those associated with contracts or sub-projects.
PMID	Unique ID assigned by PubMed to identify PubMed articles.
ProjectNumber	Project number.
SubProjectNumber	Sub project number.
AID	Unique author ID.
PI_Name	The full name of PI.
Application_ID	A unique identifier of the project record in the ExPORTER database.

22. C06_Link_Papers_BioEntities

Entity information set extracted from document titles and abstracts using BERN2.

Column Name	Description
PMID	Unique ID assigned by PubMed to identify PubMed articles.
StartPosition	Start position of mention in text.
EndPosition	End position of mention in text.
Mention	Entity mentioned in text.
Entityid(8 columns)	Normalized entity identifier, detailed in 7 columns include mesh, mim, CL, cellosaurus, NCBITaxon, NCBIGene, CHEBI.
Type	Enumerated type of entity; values include species, disease, gene, drug, mutation, cell_line, cell_type, DNA, RNA.
is_neural_normalized	For diseases and chemicals, BERN2 use hybrid NEN models, which are a combination of both rule-based and neural network-based models. An entity that is not normalized by the rule-based model is then normalized by a neural network-based model.

23. C07_Authors

Disambiguated authors in PKG.

Column Name	Description
AID	Unique author ID.
FullName	Full name of the author.
BeginYear	The year this author published the first paper.
RecentYear	The year of the author's most recent publication.
PaperNum	The number of papers published by the author.
Gender	Gender estimated based on the author's first name.
Race/Ethnic	Race/Ethnic estimated based on author's last name.
h_index	H-index of the author.

24. C08_AuthorEducations

Education information of scientific personnel from ORCID dataset.

Column Name	Description
ORCID	Unique researcher ID that distinguishes the researcher from others allocated by ORCID.
BeginYear	The beginning year of the researcher's education.
EndYear	The end year of the researcher's education.
AID	Unique author ID.
Organization	The organization the researcher has been educated.
Identifier	The identifier of an organization.
IdSource	The provider of an organizations' identifier.
Role	The degree that the researcher received.

25. C09_AuthorEmployments

Employment information of scientific personnel from ORCID dataset.

Column Name	Description
ORCID	Unique researcher ID that distinguishes the researcher from others allocated by ORCID.
BeginYear	The beginning year of the researcher's employment.
EndYear	The end year of the researcher's employment.
AID	Unique author ID.
Department	The department which the researcher belongs to.
Organization	The institution which the researcher belongs to.
Identifier	The identifier of an organization.
IdSource	The provider of an organizations' identifier.

26. C10_Link_Papers_Journals

Journal details for the year the paper was published or the most recent year.

Column Name	Description
PMID	Unique ID assigned by PubMed to identify PubMed articles.
PubYear	Year the article was published.
Journal_ISSN	Unique identifier of the journal.
Journal_RecordYear	The year of the journal information for the row.
Journal_Title	Title of the journal.
Journal_SJR	The journal's SJR for the year.
Journal_SJR_Best_Quartile	The journal's best quartile according to its SJR score the year.
Journal_Hindex	The journal's H-index for the year.
Journal_Categories	The journal's category and corresponding quartile for the year.

27. C11_ClinicalTrials

Clinical trial studies from ClinicalTrials.gov. For more information, please refer to [AACT](#) or [ClinicalTrials.gov](#).

Column Name	Description
nct_id	Unique ID assigned by ClinicalTrials.gov to identify clinical trial studies.
study_first_submitted_date	The date that the responsible party pressed the release button in the ClinicalTrials.gov data submission system.
start_date	The date the study started in a date-type format.
completion_date	The date the study was completed (as defined by NLM) in a date-type format.
official_title	The title of the clinical study, corresponding to the title of the protocol.

28. C12_Link_Papers_Clinicaltrials

Citations between PubMed papers and clinical trial studies.

Column Name	Description
PMID	Unique ID assigned by PubMed to identify PubMed articles.
nct_id	Unique ID assigned by ClinicalTrials.gov to identify clinical trial studies.
Type	Indicate the location of the reference or whether it is recorded in PubMed.

29. C13_Link_ClinicalTrials_Projects

Linkages between NIH fundings and clinical trial studies.

Column Name	Description
-------------	-------------

project_number	Unique ID assigned by NIH ExPORTER to identify NIH fundings.
nct_id	Unique ID assigned by ClinicalTrials.gov to identify clinical trial studies.

30. C14_ClinicalTrials_BioEntities

Entity information set extracted from document titles and abstracts using BERN2.

Column Name	Description
nct_id	Unique ID assigned by ClinicalTrials.gov to identify clinical trial studies.
StartPosition	Start position of mention in text.
EndPosition	End position of mention in text.
Mention	Entity mentioned in text.
Entityid(7 columns)	Normalized entity identifier, detailed in 7 columns include mesh, mim, CL, cellosaurus, NCBITaxon, NCBIGene, CHEBI.
Type	Enumerated type of entity; values include species, disease, gene, drug, mutation, cell_line, cell_type, DNA, RNA.
is_neural_normalized	For diseases and chemicals, BERN2 use hybrid NEN models, which are a combination of both rule-based and neural network-based models. An entity that is not normalized by the rule-based model is then normalized by a neural network-based model.

31. C14_Investigators

Investigators recorded in ClinicalTrials.gov, and their AID allocated in PKG.

Column Name	Description
nct_id	Unique ID assigned by ClinicalTrials.gov to identify clinical trial studies.
name	Name of the investigator.
affiliation	Affiliation text.
AID	Unique author ID.
same_author_prob	The score of the linkage between the investigator and the AID.
affiliation_disambiguated	Institution id allocated by PKG base on the affiliation.

32. C15_Patents

Patents from USPTO. For more information, please refer to PatentsView.org.

Column Name	Description
PatentId	Unique ID assigned by USPTO to identify patents.
Type	Category of patent. Usually "design", "reissue", etc.
GrantedDate	Date when patent was granted.
Title	Title of patent.

Abstract	Abstract of patent.
Kind	WIPO document kind codes.
ClaimNum	Number of claims.
isWithdrawn	Whether a patent has been withdrawn or not.
has_citing_paper	Whether a patent citing any paper.
is_granted_by_NIH	Whether a patent granted by NIH.
is_CPC_A61	Whether a patent's CPC classification is A61 (MEDICAL OR VETERINARY SCIENCE; HYGIENE).

33. C16_Link_Patents_Papers

Citations from patents to papers, it is the subset of pcs dataset run by Marx Matt. For more information, please refer to <https://zenodo.org/records/10215169>.

Column Name	Description
PatentId	Unique ID assigned by USPTO to identify patents.
WhereFound	Front only, body only, or both.
ConfScore	Assigned confidence score to the match.
SelfCitation	Indicates whether at least one author on the paper was also an inventor on the patent.
RefType	App = from applicant; Exm = from examiner; Unk = if unspecified in the unstructured reference
PMID	Unique ID assigned by PubMed to identify PubMed articles.

34. C17_Link_Patents_Projects

Linkages between NIH fundings and patents.

Column Name	Description
project_number	Unique ID assigned by NIH ExPORTER to identify NIH fundings.
PatentId	Unique ID assigned by USPTO to identify patents.

35. C18_Link_Patents_BioEntities

Entity information set extracted from document titles and abstracts using BERN2.

Column Name	Description
PatentId	Unique ID assigned by USPTO to identify patents.
StartPosition	Start position of mention in text.
EndPosition	End position of mention in text.
Mention	Entity mentioned in text.
Entityid(7 columns)	Normalized entity identifier, detailed in 7 columns include mesh, mim, CL, cellosaurus, NCBITaxon, NCBIGene, CHEBI.
Type	Enumerated type of entity; values include species, disease, gene, drug, mutation, cell_line, cell_type, DNA, RNA.
is_neural_normalized	For diseases and chemicals, BERN2 use hybrid NEN models, which are a combination of both rule-based and neural network-based models. An entity that is not normalized by the rule-based model is then normalized by a neural network-based model.

36. C19_Inventors

Inventors recorded in USPTO and PatentsView. For more information, please refer to patentsview.org.

Column Name	Description
patent_id	Unique ID assigned by USPTO to identify patents.
inventor_sequence	Order in which inventor appears in file; sequence for this table begins at {0}.
inventor_id	Unique inventor id generated by the disambiguation algorithm in PatentsView.
Disambig_inventor_name_first	First name.
disambig_inventor_name_last	Last name.

37. C20_Link_Authors_Inventors

Inventors recorded in USPTO and PatentsView, and their AID allocated in PKG.

Column Name	Description
inventor_id	Unique inventor id generated by the disambiguation algorithm in PatentsView.
AID	Unique author ID.
same_author_prob	The score of the linkage between the investigator and the AID.

38. C21_Bioentity_Relationships

Biomedical entity relationships in papers.

Column Name	Description
PMID	Unique ID assigned by PubMed to identify PubMed articles.
Entity_id1/Entity_id2	The IDs of two biomedical entities.
relation_type	Type of the two biomedical entities.
relation_id	ID allocated by PKG to identify the relation.

39. C22_DatasetMethod

Using dictionary method, extract dataset and method entities from the article.

Column Name	Description
PMID	Unique ID assigned by PubMed to identify PubMed articles.
Mention	Entity mentioned in text.
Entity	Standard expression of entity.
Type	Identify whether this entity is a method or a dataset.

40. C23_BioEntities

Information of biomedical entities.

Column Name	Description
EntityId	Normalized entity identifier.
Type	Enumerated type of entity; values include species, disease, gene, drug, mutation, cell_line, cell_type, DNA, RNA.
Mention	Entity mentioned in text.