Description About CSV Files

We built PKG with bio-entities extracted from PubMed abstracts, author name disambiguation results of PubMed authors, and the integrated multi-source information. It contains seven comma-separated value (CSV) files named "Author_List," "Bio_entities_Main," "Bio_entities_Mutation," "Affiliations," "Researcher_Employment," "Researcher_Education," and "NIH_Projects". The details are presented in Table 1. PubMed raw data are not included into CSV files because the amount of PubMed raw data is too large and they are not generated or altered by our methods. PubMed raw data can be freely downloaded from PubMed website (https://www.nlm.nih.gov/databases/download/pubmed_medline.html). We also provide the following download link (http://er.tacc.utexas.edu/datasets/ped), which contains both the PubMed raw data and PKG dataset to facilitate the application of PKG dataset.

Table 1. Dataset details.

File	**# of Lines**	**# of Distinct PMIDs**	**# of Distinct AND_IDs**	**Short description**
OA01_Author_List	131,446,038	31,270,411	18,519,492	CSV file containing PubMed authors and AND_IDs.
OA02_Bio-entities_Main	295,921,671	20,136,150	-	CSV file containing all types of extracted bio-entities by BioBERT.
OA03_Bio-entities_Mutation	1,415,427	320,025	-	CSV file containing additional items of mutations from Bio-entities_Main file.
OA04_Affiliations	62,015,712	20,941,553	9,502,394	CSV file containing affiliations and their extracted fine-grained items.
OA05_Researcher_Employment	1,194,697	-	531,916	CSV file containing employment history from ORCID.
OA06_Researcher_Education	934,507	-	448,110	CSV file containing educational background from ORCID.
OA07_NIH_Porjects	22,946,601	1,886,856	116,530	CSV file containing projects from NIH ExPORTER and mapping relation between PI_ID, PMID, and AND_ID.

Note: In file Author_List, about 2.8 million (2.16%) author instances cannot be disambiguated because they do not exist in Authority or Semantic Scholar dataset. Therefore, their AND_ID field values were set to zero.