Unlocking the Power of Data Harmonization in Environmental Health Sciences: A Comprehensive Exploration of Significance, Use Cases, and Recommendations for Standardization Efforts-Supplemental Materials

Authors

Jeanette A. Stingone,[1] HC Bledsoe,[2] Grace Cooney,[2] Mireya Diaz-Insua,[3] Elaine Faustman,[4] Karamarie Fecho,[5,6] Ramkiran Gouripeddi,[7] Philip Holmes,[8] David Kaeli,[9] Oswaldo Lozoya,[10] Anna Maria Masci,[11] Hina Narayan,[12] Charles Schmitt,[13] Maria Shatz,[13] Wren Tracy[2]

[1]Department of Epidemiology, Columbia University Mailman School of Public Health, New York, New York, USA
[2]ICF, Reston, Virginia, USA
[3]Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, Ohio, USA
[4]Department of Environmental & Occupational Health Sciences, University of Washington, Seattle, Washington, USA
[5]Copperline Professional Solutions, LLC, Pittsboro, North Carolina, USA
[6]Renaissance Computing Institute, University of North Carolina, Chapel Hill, North Carolina, USA
[7]Department of Biomedical Informatics, University of Utah, Salt Lake City, Utah, USA
[8]Department of Physics, Villanova University, Villanova, Pennsylvania, USA
[9]Department of Electrical and Computer Engineering, Northeastern University, Boston, Massachusetts, USA
[10]RTI International, Research Triangle Park, North Carolina, USA
[11]Department of Data Impact and Governance, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA
[12]Milken Institute School of Public Health, The George Washington University, Washington, District of Columbia, USA
[13]Office of Data Science, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina, USA

Address correspondence to Jeanette Stingone, Department of Epidemiology, Columbia University Mailman School of Public Health, 722 West 168th Street, New York, NY 10032 USA. Email: js5406@cumc.columbia.edu

Methodology

Comparison of different similarity models: The method deployed in the paper to map variables between two studies using computer algorithms are foremost dependent on the approach used to measure the similarity between a variable from one study and a variable from a second study. In practice, we appended the description of each variable to its variable name to form an augmented variable description. These augmented variable descriptions were then compared for similarity. Two examples of augmented variables names are: “age Child's age at baseline” and “chapsticklast Last time used”. This approach ensures that the information available in both the variable name and the description are available to the similarity algorithm. As illustrated with “chapsticklast Last time used” the variable name “chapsticklast” includes the important aspect of the product chapstick and the description includes the important concept of time.

We tested four different approaches to measure similarity to better understand the impact on matching human mapping results. These are:

3large: in this approach, each augmented variable description is converted to a mathematical embedding vector of length 3072 using the OpenAI large language based model “text-embedding-3-large”. The similarity between two variables is then computed as the cosine similarity of their mathematical vectors. The 3large model is the highest performance of the OpenAI embedding models (based on OpenAI benchmarks) available at the time this work was conducted.

3small: identical to the 3large approach, but using the OpenAI model text-embedding-3-small”, a lower price and lower performance model that uses embeddings of length 1536.

Ada002: identical to the 3large approach, but using the OpenAI model text-embedding-ada-002”, a model that uses embeddings of length 1536 that was developed prior to the text-embedding-3-small and large models, is lower price, and has lower performance on OpenAI benchmarks than the other OpenAI models.

Tfidf (Term Frequency - Inverse Document Frequency): Tfi-df is a methodology created prior to the creation of large language models that allows for comparing two or more sets of documents while accounting for the frequency of words (to lower the importance of matching common words such as ‘the’). This was included to help understand how the introduction of large language technologies has advanced the task of mapping variables between studies.

Results

The comparison of approaches is provided in the table below. The 3large model outperformed the other models based on F1 score for two of the study pairs (ie Study A x Study B, Study C x Study B). 3small outperformed 3large for one study pair (Study D x Study E) although the difference is small (0.78 vs 0.76). Interestingly, tfidf outperformed the large language models approaches in one case (Study C x Study D) by a considerable amount. Overall, however, it’s clear that large language models have advanced performance. While there is considerable room for improvement, the performance of the 3large model is high enough that computer based mappings could be used to assist human mapping efforts.

Table: Results from comparison of different similarity models

MODEL	STUDY1	STUDY2	THRESHOLD	RECALL	PRECISION	F1	TP	FP	FN
3LARGE	**Study A**	**Study B**	**0.45**	**0.26**	**0.60**	**0.36**	**12**	**34**	**8**
3SMALL	Study A	Study B	0.40	0.20	0.45	0.28	9	36	11
TFIDF	Study A	Study B	0.25	0.75	0.15	0.25	3	1	17
ADA002	Study A	Study B	0.80	0.05	0.50	0.10	10	178	10
3SMALL	**Study E**	**Study D**	**0.65**	**0.75**	**0.82**	**0.78**	**9**	**3**	**2**
3LARGE	Study E	Study D	0.65	0.80	0.73	0.76	8	2	3
TFIDF	Study E	Study D	0.35	0.88	0.64	0.74	7	1	4
ADA002	Study E	Study D	0.80	0.09	1.00	0.17	11	109	0
TFIDF	**Study C**	**Study D**	**0.30**	**0.69**	**1.00**	**0.82**	**9**	**4**	**0**
3LARGE	Study C	Study D	0.65	0.67	0.67	0.67	6	3	3
3SMALL	Study C	Study D	0.65	0.63	0.56	0.59	5	3	4
ADA002	Study C	Study D	0.80	0.06	1.00	0.11	9	140	0
3LARGE	**Study C**	**Study E**	**0.55**	**0.56**	**0.82**	**0.67**	**14**	**11**	**3**
3SMALL	Study C	Study E	0.60	0.55	0.65	0.59	11	9	6
TFIDF	Study C	Study E	0.20	0.31	0.71	0.43	12	27	5
ADA002	Study C	Study E	0.80	0.07	1.00	0.13	17	225	0

Bolded rows indicated the highest performance model for each pair of studies evaluated. See paper for definition of columns.