A Unified Rubric and Similarity Metric for Biomedical Publication Types and Study Designs

*Neil R. Smalheiser[1], MD, PhD, Joe D. Menke[2], MS and Arthur W. Holt[1], MBA

*Corresponding author: Department of Psychiatry MC912, 1601 W. Taylor St., Chicago, IL 60612. neils@uic.edu, 708-369-3591

[1]Department of Psychiatry, University of Illinois Chicago, Chicago, IL

[2]School of Information Sciences, University of Illinois Urbana-Champaign, Champaign, IL

Keywords: Classification; Information Storage and Retrieval; MEDLINE; PubMed; Databases, Bibliographic

Word count, excluding title page, abstract, references, figures and tables: 2195

Abstract

Objective. Our goal is to unify the hierarchies of 72 biomedical publication types and study designs (collectively, PTs) into a single rubric. **Materials and Methods.** This is carried out in a data-driven manner by computing pairwise similarities of each PT against all others to form a similarity matrix. By performing hierarchical clustering we place each PT in a specific category and collect these into broader categories. **Results.** Spearman correlations among PT pairs ranged from strongly negative to strongly positive (-0.732 to +0.997), with a mean of 0.176. Overall, we obtained 13 clusters of PTs and 5 more general categories: Observational Clinical Research, Qualitative and Genetic Methods, Clinical Evaluation and Validation, Interventional Trial

Research, and Scholarly Synthesis and Discourse. **Discussion.** The rubric provides a flexible classification scheme for publication types and study designs that can accommodate new PTs as they are added over time. **Conclusion.** The PT rubric provides an overview that complements the existing NIH hierarchies. Further, the similarity metric has the potential to improve the modeling and evaluation of automated indexing systems.

Background and Significance

Arguably the best curated indexing scheme for biomedical articles is the Medical Subject Heading (MeSH) terminology developed at the National Library of Medicine (NLM).[1] Each article in MEDLINE is annotated with a set of 8-20 terms, according to the most important topics discussed. The terminology is arranged first in categories and then placed in hierarchies, according to their logical relationships. A separate hierarchy deals with Publication Types (PT) such as Review, Letter, or Editorial (https://www.nlm.nih.gov/mesh/pubtypes.html). Until recently all indexing was performed by PhD-level curators who read the full-text of each article and assign the most specific applicable term(s) on the hierarchy.[2] Currently, however, most indexing is carried out by automated machine learning systems.[3] The scheme is designed to facilitate high-recall retrieval of relevant articles when a user enters a free-text search query into the PubMed search interface. In the PubMed back-end, queries are processed to identify relevant MeSH terms and searches are automatically expanded to include synonyms and more specific terms that lie under them in the hierarchy.[4]

In contrast to indexing of MeSH headings for the most important topics discussed in an article, some unique issues relate to automated indexing of articles in terms of publication types and

study designs (for example, case-control studies, cross-over studies, etc.).[5-8] Our research team has been engaged in a long-term project along these lines.[9-13] We seek to tag each article explicitly with all indexing terms that apply to them. In addition, we recognize that some articles are atypical; thus, we do not simply generate binary yes/no predictions but also assign probabilistic scores for each term. Together, this strategy is designed to ensure that a user seeking to retrieve all articles of a given study design can identify them with high precision and recall, and further, can adjust the threshold of the probabilistic scores to either maximize precision or recall as desired.[14, 15] For simplicity, both publication types (PT terms) and study design-related MeSH terms will be collectively referred to as PTs.

During the course of our indexing project, we have identified several limitations regarding the NIH MeSH and PT hierarchies. First, the Publication Types form a separate set of trees apart from the study designs that are listed on MeSH trees. Yet many Publication Types imply particular study designs; for example, Randomized Controlled Trial is listed as a Publication Type but has definite design features (random allocation, comparing an experimental group against a control group, generally interventional, and often double-blind). Thus, it would be desirable to merge publication types and study design-related MeSH terms into a single rubric. Second, the relative position of two terms on the hierarchical trees do not necessarily provide a measure of how similar they are. A similarity metric would provide a fuller understanding of how PTs relate to each other.

Objective

Our goal is to unify the hierarchies of publication types and study designs into a single rubric. This is carried out in a data-driven manner by computing pairwise similarities of each PT against all others to form a similarity matrix. By performing hierarchical clustering we place each PT in a specific category and collect these into broader categories.

Materials and Methods

In a previous publication type indexing study [11] we processed the title and abstract of each article in the biomedical literature to create a vector representation of the article, then collected the vector representations of all articles belonging to a given PT and designated the centroid vector to represent the PT as a whole. In such a scheme, the similarity of two PTs can be measured in a straightforward manner as the distance between their centroids [16]. This metric was employed in a SVM-based predictive model of 50 PTs called Multi-Tagger.[11] However, more recently, we have utilized a transformer-based model that employed BERT-based encoders and represents articles not only in terms of text but some non-textual features as well (e.g., journal name, number of authors, number of references) [12, 13], and infers a probability value ($0<p<1$) that a given article belongs to each of 72 PTs. This model was employed here since it is more comprehensive and has been shown to have higher overall predictive performance for indexing PTs than the previous SVM-based model.[12]

Correlation analysis

The model probabilities produced by the transformer model were employed to obtain a pairwise similarity metric across all PTs. Full details of the model, its training data, and its architecture are given in Supplemental File 1. Briefly, the transformer model uses WeighCon supervised

contrastive learning, and title, abstract, and various metadata-based features, such as journal name, served as input into the model. Training examples were taken from PubMed records published between 1987 and 2023 that had been manually assigned PTs by NLM, adjusting the number of examples per PT by under-sampling to deal with the fact that certain PTs have many more articles than others [12]. In addition, using this same transformer model architecture, we added a new PT "Case Series", designated from a curated corpus [17], combined with the data for the other PTs. Overall, this model was developed using a stratified random sample of 1,284,378 articles, which was stratified by PT to preserve label distribution. The dataset was split using 70/10/20 train/validation/test splits, again stratified to preserve label distribution in each split. Articles were tagged if the probabilistic prediction of the model was above a threshold, which was determined empirically to optimize for F1 within the validation set. The test set of articles from the modeling work, containing 256,897 articles, formed the basis for the present analysis and was further filtered to include only articles with complete abstracts, leaving 210,641 articles for analysis. (Articles lacking abstracts were excluded because they have limited textual features and on average show lesser predictive accuracy.[12])

From the model-predicted probabilities for each of the 72 publication types across the articles of the test set, a nonlinear (Spearman) correlation rho was computed for all 2,556 pairwise combinations of the 72 publication types (illustrated in Figure 1) to form a 72x72 correlation matrix (Supplemental File 2).

Figure 1. Correlation Matrix Derivation

Model Predictions per Article

PMID	Case-Control Studies	Case Reports	Case Series	Clinical Study	...
1279786	0.0159	0.0023	0.0003	0.9656	...
1281066	0.0079	0.0103	0.0005	0.2203	...
1281180	0.0088	0.7842	0.0009	0.9575	...
1281728	0.0008	0.0006	0.0016	0.9444	...
1282546	0.9175	0.0226	0.0005	0.9620	...
1283031	0.0005	0.0051	0.0002	0.1194	...
1283035	0.0100	0.0041	0.0008	0.9334	...
1283061	0.0015	0.0079	0.0002	0.8994	...
1283513	0.0005	0.0110	0.0007	0.9639	...
⋮	⋮	⋮	⋮	⋮	

Spearman Correlation Matrix of Modeled Probabilities

	Case-Control Studies	Case Reports	Case Series	Clinical Study	...
Case-Control Studies	1	0.1572	-0.0001	0.1022	...
Case Reports	0.1572	1	0.1989	-0.1985	...
Case Series	-0.0001	0.1989	1	0.0378	...
Clinical Study	0.1022	-0.1985	0.0378	1	...
⋮	⋮	⋮	⋮	⋮	

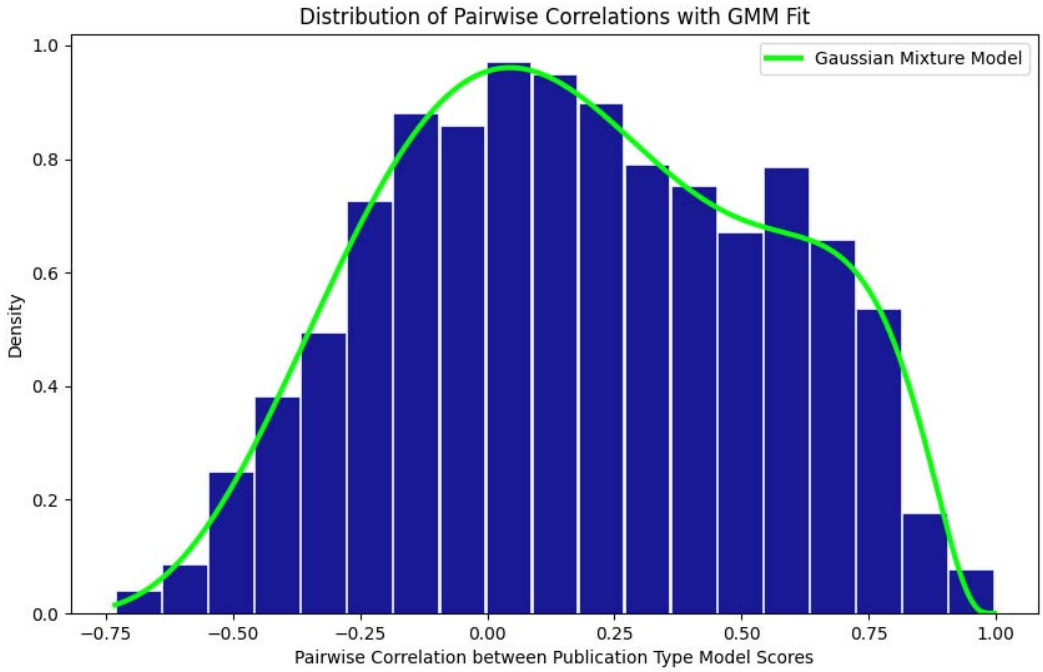
Modeled predicted probabilities were used to compute Spearman correlations for each pairwise combination of publication types.

Results

Distribution of PT Similarity Measures

Model score correlations among PT pairs ranged from strongly negative to strongly positive (-0.732 to +0.997), with a mean of 0.176. The distribution appears to be bimodal with one mode centered near zero correlation, and a shoulder centered around 0.59 (Figure 2). The highest positive correlations relate to clinical trial subtypes and related clinical study methodologies. The negative correlations consist of publication type pairs that are functionally unrelated, such as Clinical Study vs. Biography.

Figure 2. Distribution of Pairwise Correlations for PTs

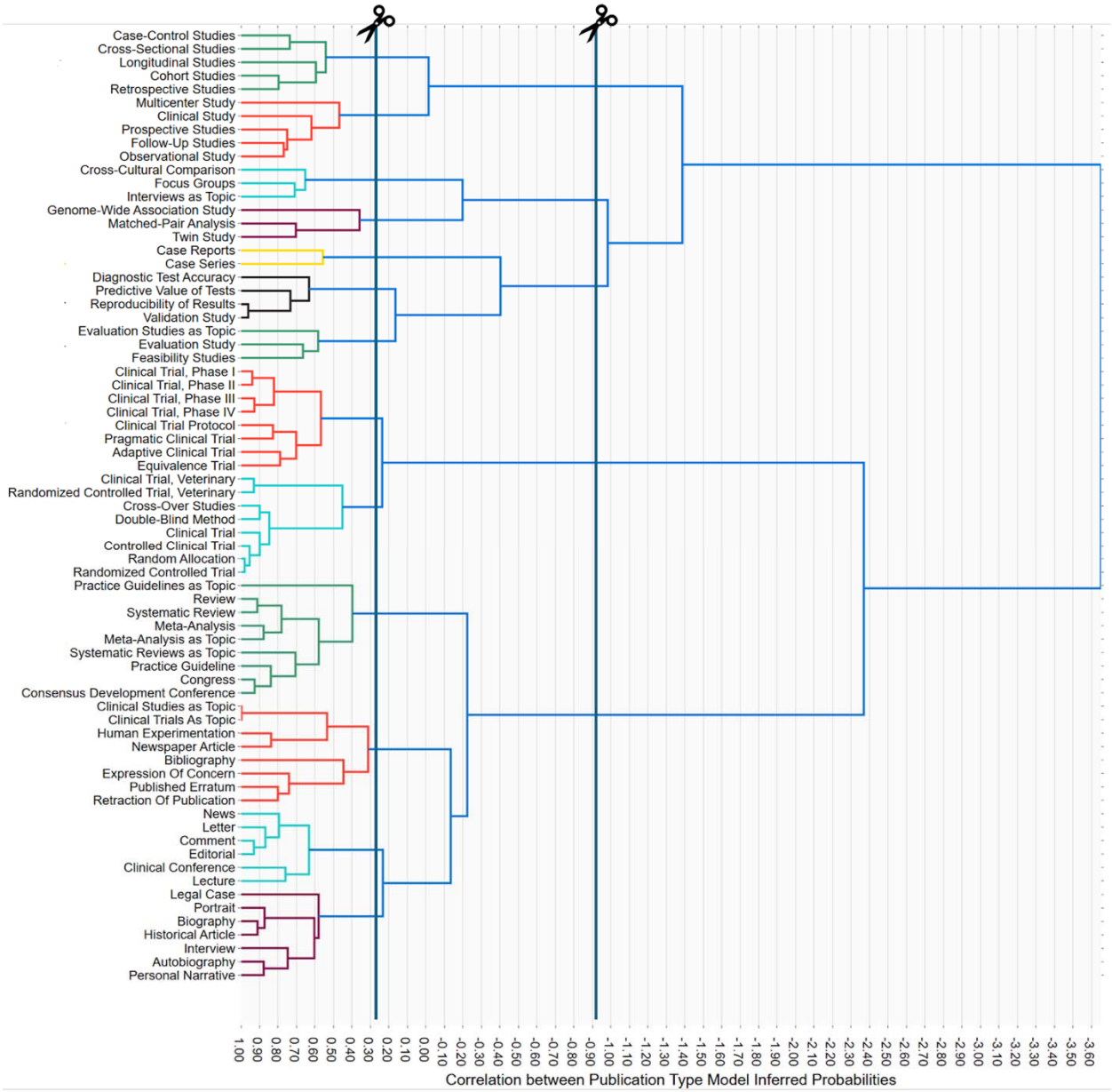


Spearman correlation was computed among model-predicted probabilities for assignment of 72 publication types forming a 72x72 correlation matrix. The distribution of the correlation coefficients is bimodal with one mode centered around zero, the other around 0.57.

Hierarchical clustering of publication types and study designs

Several different clustering techniques were evaluated, including graph-based community detection methods and statistical clustering methods. The graph-based clustering methods produced reasonable communities, but were found to be sensitive to noise and joined disparate publication types such as case reports with editorials and commentary. Agglomerative hierarchical clustering provided a more reasonable and functional basis for grouping PTs using the pairwise correlation matrix inverted to represent distances and joined with Ward linkage (Figure 3).[18]

Figure 3. Hierarchical Clustering of Publication Types by Model-Inferred Probabilities



Publication type clusters derived from pairwise model score correlations and grouped with hierarchical clustering. Cut points were chosen to delineate 13 low-level categories and 5 broader categories.

Starting with the most similar pairs of PTs, several different categories naturally emerged in a data-driven manner (Figure 3). For example, the category of Observational Epidemiologic Study Designs comprised case-control, cross-sectional, longitudinal, cohort and retrospective studies. Interestingly, other types of observational studies fell into a distinct category, General Clinical

and Observational Studies, consisting of both publication types and MeSH study designs, namely, multicenter study, clinical study, prospective studies, follow-up studies and observational study.

Note that the clustering scheme reflects the behavior of NLM indexers, who attach only the most specific indexing term that applies to a given article. Thus, even though the Clinical Study PT definition comprises both interventional and observational designs (https://www.ncbi.nlm.nih.gov/mesh/2009830), if the article is interventional, NLM indexers would affix the more specific term Clinical Trial instead. For this reason, in Figure 3, Clinical Study is placed with observational designs, whereas Clinical Trial (logically a subset of Clinical Study) is placed with interventional designs. Similarly, if an observational study had, say, a cohort design, it would be indexed as Cohort Studies instead of Observational Study, explaining why the two PTs were placed in distinct clusters. The relationships shown in Figure 3 are thus complementary to the hierarchical relationships and trees displayed in the NIH Hierarchies.

Overall, we obtained 13 low-level clusters of PTs and 5 more general categories: Observational Clinical Research, Qualitative and Genetic Methods, Clinical Evaluation and Validation, Interventional Trial Research, and Scholarly Synthesis and Discourse. One PT, Scientific Integrity Review, was excluded from the clustering process because very few articles in our dataset were tagged with that type; due to its nature (consisting of official scientific misconduct findings), it was manually placed in Scholarly Publishing and Research Integrity. Table 1 contains a full list of publication types and study designs examined along with category assignments. Another view is presented in Supplemental File 2.

Table 1. Publication Types and Observed Cluster Assignments

Publication Type	Low Level Category	Broader Category
Case-Control Studies Cross-Sectional Studies Longitudinal Studies Cohort Studies Retrospective Studies	Observational Epidemiologic Study Designs	Observational Clinical Research
Multicenter Study Clinical Study Prospective Studies Follow-Up Studies Observational Study	General Clinical & Observational Studies	Observational Clinical Research
Cross-Cultural Comparison Focus Groups Interviews as Topic	Qualitative & Sociocultural Research Methods	Qualitative & Genetic Methods
Genome-Wide Association Study Matched-Pair Analysis Twin Study	Genetic & Matched Population Analyses	Qualitative & Genetic Methods

Case Reports **Case Series**	**Clinical Case-Based Evidence**	**Clinical Evaluation & Validation**
Diagnostic Test Accuracy **Predictive Value of Tests** **Reproducibility of Results** **Validation Study**	**Diagnostic & Methodological Validation Studies**	**Clinical Evaluation & Validation**
Evaluation Studies as Topic **Evaluation Study** **Feasibility Studies**	**Program & Process Evaluation Studies**	**Clinical Evaluation & Validation**
Clinical Trial, Phase I **Clinical Trial, Phase II** **Clinical Trial, Phase III** **Clinical Trial, Phase IV** **Clinical Trial Protocol** **Pragmatic Clinical Trial** **Adaptive Clinical Trial** **Equivalence Trial**	**Interventional Clinical Trial Phases & Designs**	**Interventional Trial Research**

Clinical Trial, Veterinary **Randomized Controlled Trial, Veterinary** **Cross-Over Studies** **Double-Blind Method** **Clinical Trial** **Controlled Clinical Trial** **Random Allocation** **Randomized Controlled Trial**	**Controlled & Randomized Trial Methodology**	**Interventional Trial Research**
Practice Guidelines as Topic **Review** **Systematic Review** **Meta-Analysis** **Meta-Analysis as Topic** **Systematic Reviews as Topic** **Practice Guideline** **Congress** **Consensus Development Conference**	**Evidence Synthesis & Clinical Guidance**	**Scholarly Discourse and Evidence Synthesis**

News **Letter** **Comment** **Editorial** **Clinical Conference** **Lecture**	**Scientific Commentary & Professional Discourse**	**Scholarly Discourse and Evidence Synthesis**
Legal Case **Portrait** **Biography** **Historical Article** **Interview** **Autobiography** **Personal Narrative**	**Biographical, Historical & Narrative Works**	**Scholarly Discourse and Evidence Synthesis**

Discussion

The formulation of a similarity metric for PTs led to the construction of a unified rubric that provides a simple, concrete overview of the many biomedical publication types and study designs, in a single diagram. This provides a complementary view from the NIH hierarchies which identify different terms in terms of their logical relationships and proximity on several independent trees. Cid and Mork previously constructed a PT correlation matrix based on the frequency of co-occurrence of multiple PTs appearing on individual articles,[8] but their scheme did not include study design-related MeSH terms and only considers articles that received

multiple PT terms. In contrast, our similarity metric was computed across all articles and reflects the indexing decisions of NLM indexers, who generally attach only the most specific indexing term that applies to a given article. Hence our scheme provides a more comprehensive, finer-grained overview.

As the similarity metric is data-driven, it can be recalculated to adapt to future changes in publications, and can easily accommodate new indexing terms as they arise over time. For example, since NLM currently has no formal indexing of case series articles, we recently created a new publication type, Case Series[17]. After adding this PT to the existing transformer model we obtained correlation values which placed Case Series next to Case Reports (Figure 2).

Apart from providing a classification scheme for publication types and study designs, we believe that the similarity metric should have utility for better automated PT indexing. With regard to the current transformer model, we use a one-against-all model in which each article is assessed for a given PT vs. all others. However, the training process should be sharpened if one used training examples of one PT as negative examples against another PT that is very similar (but not a subset of the other). For example, Case Reports and Case Series are nearest neighbor PTs in the rubric. Case reports tend to be incidental observations of one or a few patients, whereas case series tend to be planned observations of four or more patients, but the two types of article are not sharply delineated.[17] By comparing and contrasting these two PTs during model training, the probabilistic scores should better discriminate whether a given article has typical case report-like vs. typical case series-like features.

The similarity metric may also assist in evaluating the errors made by an automated indexing system. Since the probabilistic predictive scores for nearest-neighbor PTs have high Spearman

correlation values, one would expect that a mis-assigned article would most likely be tagged with the nearest neighbor PT instead. In many cases, the errors might be of little import (e.g., mis-assigning a Focus Group article as Interviews as Topic). Other nearest-neighbor errors have high importance for high recall retrieval, for example, mis-assigning Cohort Studies as Retrospective Studies. Again, identifying these cases would prioritize them for contrastive training in which one PT is used as negative examples for the other PT.

Another possible use for the rubric is as a framework for alternative LLM-based assignment of publication types and study designs.[6] For example, the organization of PTs into small and larger categories allows one to envision a Twenty Questions procedure in which the LLM would first attempt to place an article into the most appropriate broad category (e.g. observational vs. interventional research) and then to ascertain the best specific category and the best choice among the PTs in that category.

This study has several limitations. The processing of articles was limited to articles containing abstracts, which provide more complete data for model predictive scores; this will under-sample a few PTs that have a high proportion of articles lacking abstracts (e.g., Letter or Portrait). However, these PTs tend to be the least important in terms of clinical relevance and should have little effect on the rubric as a whole. Another limitation is that agglomerative hierarchical clustering's optimal solution is not necessarily unique. However, although calculated nearest-neighbors can vary slightly across different test sets, different model parameters, and different examined time periods, the placement of PTs into categories is likely to be robust and stable over time.

Conclusion

We present here pairwise similarity measures that allowed us to construct a rubric that arranges publication types and study designs in a unified diagram. Such a rubric complements the existing NIH hierarchies and has the potential to improve the modeling and evaluation of automated indexing systems.

Conflict of Interest statement

The authors declare that they have no competing interests.

Funding

This work was supported by the National Library of Medicine at the National Institutes of Health [1R01LM014292-01 to N.R.S.]. Funder had no influence on the study, its design, or its publication.

Data Availability

The nonlinear (Spearman) correlation rho was computed for all pairwise combinations of the 72 publication types (illustrated in Figure 1) to form a 72x72 correlation matrix. This is shown in Supplemental File 2.

Author Contributions

AWH: Methodology, Formal analysis, Investigation, Writing – review & editing.

JDM: Methodology, Software, Validation.

NRS: Conceptualization, Funding acquisition, Methodology, Supervision, Writing - original draft, Writing – review & editing.

References

1. 50th Anniversary Medical Subject Headings (MeSH) Event. Technical Report. Available online at: https://www.nlm.nih.gov/pubs/techbull/nd10/nd10_mesh_50th_anniv.html

2. Rae AR, Mork JG, Demner□Fushman D. The National Library of Medicine indexer assignment dataset: A new large□scale dataset for reviewer assignment research. Journal of the Association for Information Science and Technology. 2023 Feb;74(2):205-18.

3. Krithara A, Mork JG, Nentidis A, Paliouras G. The road from manual to automatic semantic indexing of biomedical literature: a 10 years journey. Front Res Metr Anal. 2023 Sep 29;8:1250930. doi: 10.3389/frma.2023.1250930.

4. National Library of Medicine. [2018-10-29]. How PubMed works: Automatic Term Mapping https://www.nlm.nih.gov/bsd/disted/pubmedtutorial/020_040.html

5. Alhelou C, Oketola P, Allam N, Arbona JF, Clement R, Dufour J. Training an AI Model for Publication Type Classification: A Case Study on Alopecia Areata. Journal of Medical Artificial Intelligence. 2025 May 26. https://jmai.amegroups.org/article/view/10187

6. Doneva SE, de Viragh S, Hubarava H, Schandelmaier S, Briel M, Ineichen BV. Studytypeteller—large language models to automatically classify research study types for systematic reviews. Research Synthesis Methods. 2025 Sep 11:1-20.

7. Maisano DA, Ferrara L, Franceschini F. Errors in document-type classification: a focus on engineering publications and their publishers. Scientometrics. 2025 Oct 16:1-29.

8. Cid VH, Mork J. Enhancing Automatic PT Tagging for MEDLINE Citations Using Transformer-Based Models. ArXiv [Preprint]. 2025 Jun 3:arXiv:2506.03321v1.

9. Cohen AM, Smalheiser NR, McDonagh MS, Yu C, Adams CE, Davis JM, Yu PS. Automated confidence ranked classification of randomized controlled trial articles: an aid to evidence-based medicine. Journal of the American Medical Informatics Association. 2015 May 1;22(3):707-1.

10. Cohen AM, Dunivin ZO, Smalheiser NR. A probabilistic automated tagger to identify human-related publications. Database. 2018;2018:bay079.

11. Cohen AM, Schneider J, Fu Y, McDonagh MS, Das P, Holt AW, Smalheiser NR. Fifty ways to tag your pubtypes: Multi-tagger, a set of probabilistic publication type and study design

taggers to support biomedical indexing and evidence-based medicine. medRxiv. 2021 Jul 16:2021-07. https://www.medrxiv.org/content/10.1101/2021.07.13.21260468v1

12. Menke JD, Kilicoglu H, Smalheiser NR. Publication Type Tagging using Transformer Models and Multi-Label Classification. AMIA Annu Symp Proc. 2025 May 22;2024:818-827.

13. Menke JD, Ming S, Radhakrishna S, Kilicoglu H, Smalheiser NR. Enhancing automated indexing of publication types and study designs in biomedical literature using full-text features. medRxiv [Preprint]. 2025 Apr 28:2025.04.23.25326300. doi: 10.1101/2025.04.23.25326300.

14. Schneider J, Hoang L, Kansara Y, Cohen AM, Smalheiser NR. Evaluation of publication type tagging as a strategy to screen randomized controlled trial articles in preparing systematic reviews. JAMIA Open. 2022 Mar 30;5(1):ooac015. doi: 10.1093/jamiaopen/ooac015.

15. Proescholdt R, Hsiao TK, Schneider J, Cohen AM, McDonagh MS, Smalheiser NR. Testing a filtering strategy for systematic reviews: evaluating work savings and recall. AMIA Jt Summits Transl Sci Proc. 2022 May 23;2022:406-413.

16. Smalheiser NR, Cohen AM. Design of a generic, open platform for machine learning-assisted indexing and clustering of articles in PubMed, a biomedical bibliographic database. Data Inf Manag. 2018 Jun;2(1):27-36. doi: 10.2478/dim-2018-0004.

17. Shahidehpour A, Holt AW, Troy AM, Menke JD, Smalheiser NR. Creating an Indexing Scheme for Case Series Articles. medRxiv. 2025:2025-12. https://www.medrxiv.org/content/10.64898/2025.12.19.25342712v1

18. Nielsen, F. (2016). Hierarchical Clustering. In: Introduction to HPC with MPI for Data Science. Undergraduate Topics in Computer Science. Springer, Cham. https://doi.org/10.1007/978-3-319-21903-5_8.

19. Singh, A., D'Arcy, M., Cohan, A., Downey, D., & Feldman, S. (2022). SciRepEval: A Multi-Format Benchmark for Scientific Document Representations. *Conference on Empirical Methods in Natural Language Processing*.

20. Liu, Liyuan, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the Variance of the Adaptive Learning Rate and Beyond. In *8th International Conference on Learning Representations, ICLR 2020, April 26-30, 2020*. OpenReview.net, 2020. https://openreview.net/forum?id=rkgz2aEKDr.

Supplemental File 1. Description of the transformer model used to generate model probabilities for publication types and study designs.

Supplemental File 2. Publication type details, including low level and broader category assignments, are provided in an Excel workbook containing two tabs as follows:

Unified Rubric - a table with 73 rows, including header row, with the following columns:

- **Publication Type** - names of publication types/study designs with capitalization designed to match NLM spelling
- **Low Level Category Number** - an integer assigned to each of the 13 low level categories described in Table 1
- **Display Sort Order** - an integer indicating position in the dendrogram presented in Figure 2
- **Low Level Category** - descriptive name for the 13 categories to which the PT is assigned
- **Broader Category** - descriptive name for the 5 broader categories to which the PT is assigned
- **Top Correlated PT to 5th Correlated PT** - the top 5 most correlated PTs for each PT, along with Spearman correlation values
- **Least Correlated PT & Least Correlation** - the PT having the least (most negative) correlation value

Correlation Matrix - a 72x72 square matrix containing the computed pairwise correlations among the 72 PTs examined (as in Figure 1)