PubMed tags published articles with greater accuracy than Embase. Other biomedical databases provide insufficient data to enable comparative analysis

Joseph Cutteridge,[1] Kim Wager[2]

[1]York and Scarborough Teaching Hospitals NHS Foundation Trust, UK
[2]Oxford PharmaGenesis, Oxford, UK

Accuracy of article tagging across major biomedical databases: a pilot study

Background

- Many literature databases tag articles to aid categorization of publications.
- However, the accuracy of article tagging in major biomedical databases remains unknown.

Objective

- To assess the accuracy of article tagging across five major biomedical databases.

Research design and methods

- In this pilot study, data were extracted from five major biomedical databases: PubMed, Embase, Dimensions, OpenAlex and BASE, using the search term 'neuronal ceroid lipofuscinoses' (**Figure 1**).

Figure 1. Data were extracted from five databases and manually annotated. Article tags provided by the databases were then compared with the manual annotations.

Results

- BASE data were not analysed because no digital object identifier was available, preventing accurate data unification.
- Dimensions and OpenAlex tags only differentiated between articles and book chapters, not article type, so were not analysed further.
- Of the 1281 articles analysed, PubMed tags matched the ground truth in 92.6% of cases, versus 68.5% for Embase (**Figure 2**).

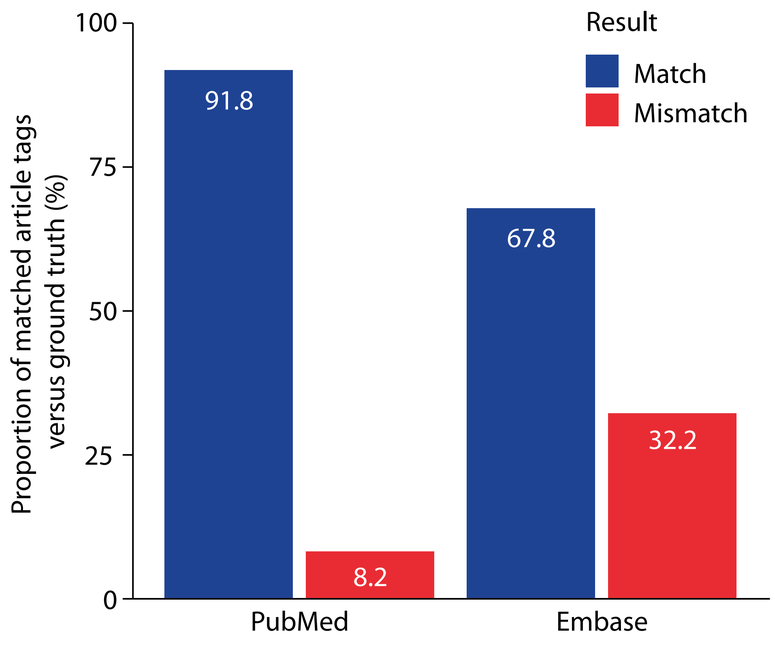

Figure 2. PubMed article tagging was more accurate than Embase article tagging.

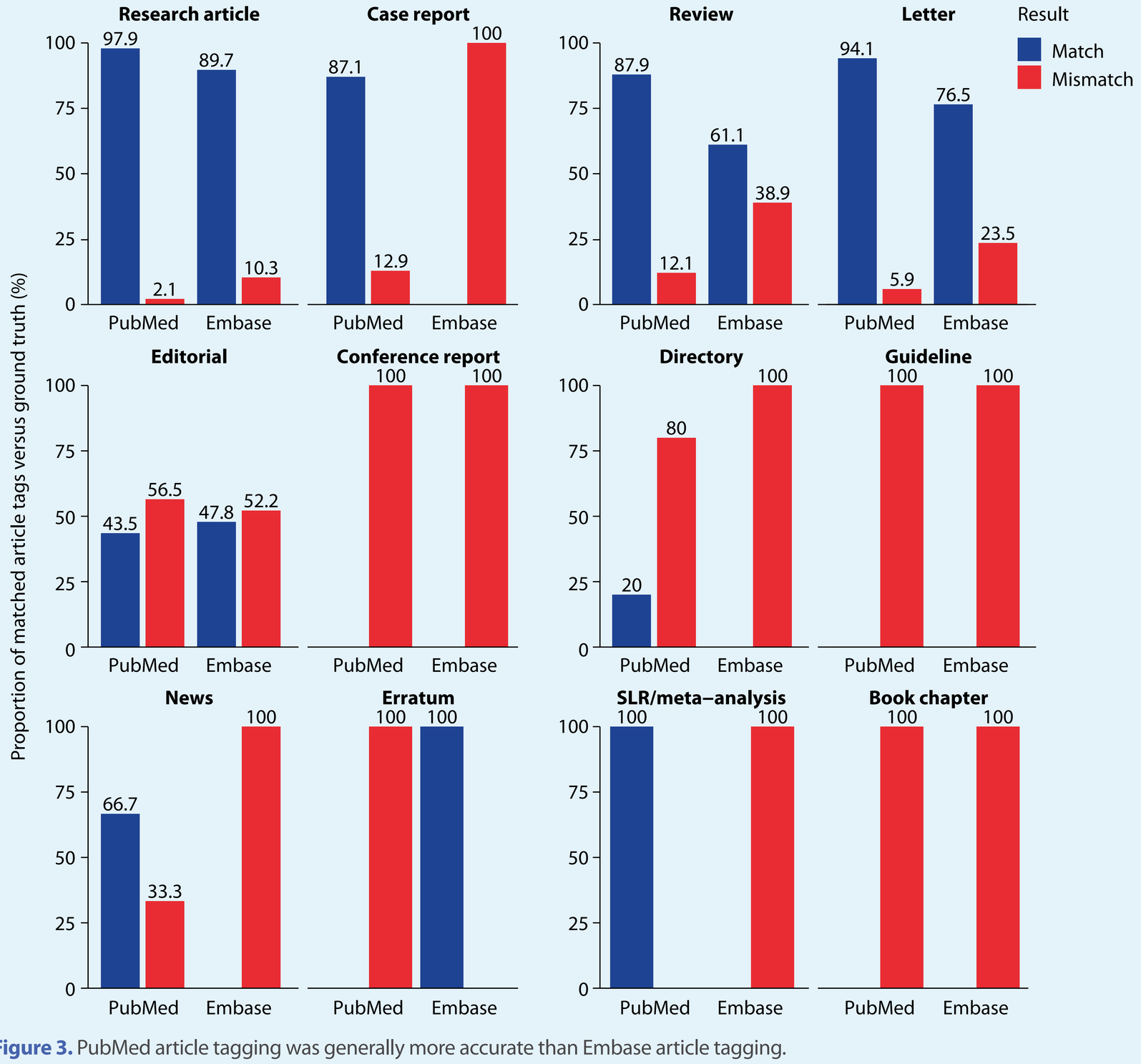

Figure 3. PubMed article tagging was generally more accurate than Embase article tagging. Embase correctly tagged conference reports more often than PubMed.
SLR, systematic literature review.

- Article tagging accuracy varied greatly by article type; however, PubMed article tags were generally more accurate than Embase tags, except for conference reports (**Figure 3**).
  - The number of articles for 'Directory', 'Guidelines', 'News', 'Erratum', 'SLR/meta analysis' and 'Book chapter' was low and so little can be inferred from the result.

Limitations

- Article categories were pre-selected to enable direct comparisons between databases. For example, inclusion of case reports heavily influenced the study outcome.
  - Assignment of ground-truth tags was provided by a single analyst, so labelling errors may have been present.

Further information

Tagging accuracy

- Embase tagging was less accurate than PubMed tagging, with two key factors accounting for 76.1% of all Embase mismatches.
  1. Embase did not tag case reports as entities distinct from research articles (n = 203).
  2. Embase incorrectly tagged many full-length research articles as conference reports (n = 103).
- PubMed performed much better at tagging review articles than Embase (87.9% vs 61.1%, respectively).

Tagging approach

- PubMed metadata are enriched by the use of multiple tags, when applicable (e.g. 'Research article', 'Randomised Control Trial', 'US Gov Funded').
- In contrast, Embase uses a single best tag approach.
- In this study, to allow direct comparison with Embase, the most specific PubMed tag was used.

Conclusions

- Accurate article tagging is important for database users.
  - Synthetic research often requires a specific article type (e.g. reports of randomized controlled trials).
  - Clinicians may only seek review articles to gain a high-level understanding of a new area.
- The broader implication is that researchers, clinicians and patients all rely on improved transparency, inclusivity and discoverability of articles, which is in part reliant on accurate article metadata.

Author contributions

Substantial contributions to study conception/design, or acquisition/analysis/interpretation of data: JC, KW.
Drafting of the publication or revising it critically for important intellectual content: JC, KW.
Final approval of the publication: JC, KW.

Disclosures

KW: Employee of Oxford PharmaGenesis.

Presented at the European Meeting of ISMPP | 24–25 January 2023