International Journal of Population Data Science

Journal Website: www.ijpds.org

Mining academic publications to automatically identify data sources

Anastasiou, A[1*] and Tingay, K[2]

[1]bibInsight, Swansea University Medical School
[2]ADRC-W, Swansea University Medical School

Background

Discovering suitable datasets is an important part of health research, particularly for projects working with cohort data, but with the proliferation of so many national and international initiatives, it is becoming increasingly difficult for research teams to locate real world datasets that are most relevant to their project objectives.

Methods

To assist researchers in this, we developed bibInsight, a data analysis platform to identify potentially useful data sources and more generally enable large scale research over bibliographical datasets.

Data source names were identified from a broad, topic-specific literature search. Context-specific terms like "annual", "longitudinal", and "prospective" were used to train a classifier that identified potential datasets.

Results

The classifier was able to identify 1588 of 1961 abstracts as containing cohort-relevant information: a precision of approximately 80%.

Further analysis such as topic analysis, geographical mapping, and collaboration networks can refine and prioritise the search results to determine the most relevant data source(s) for a research project.

Conclusions

A very large amount of information, including data source description and use, remains unexploited in unstructured bibliographical datasets. Here, we used a thematic search to provide a more manageable starting point towards locating disease specific datasets.

*Corresponding Author:
Email Address: A.Anastasiou@swansea.ac.uk (A Anastasiou)

http://dx.doi.org/10.23889/ijpds.v3i2.532