A TEXT ANALYTICS APPROACH TO SUPPORT AOP DEVELOPMENT

Identifying Scientific Literature to Support Hypothesized Adverse Outcome Pathways

Robyn Blain, Anna Engstrom, Arun Varghese | ICF

Dream big.
Then call ICF.

Contact
Arun Varghese
arun.varghese@icf.com

ICF

The Role of Literature Search in AOP Development

What are AOPs? Adverse Outcome Pathways are conceptual frameworks that evaluate how biological perturbations (molecular initiating events or MIEs) generate impacts (key events) through the molecular, cellular, and organ levels towards an adverse outcome relevant to risk assessment.

How are AOPs Developed? One commonly used approach for developing AOPs is the "top-down," hypothesis-driven approach, which starts with an experimentally determined adverse outcome and then identifies an MIE and a sequence of key events (supported by weight-of-evidence) to explain the observed outcome. This approach is resource intensive because it is dependent on experimental *in vivo* toxicology.

Alternatively, a "bottom-up" approach can be used, in which an MIE and key events are identified and then computational models, *in silico*, and *in vitro* data are used to predict a potential adverse outcome. This approach has the potential to predict spurious or irrelevant adverse outcomes.

Why is Literature Search Important? There is already a significant amount of mechanistic data available in the peer-reviewed literature. Thus, conducting a literature review is an important part of validating bottom-up approaches to AOP development.

How Can Text Analytics Help? Text analytics uses mathematical representations of language to classify and cluster documents, and to find lexically similar words and phrases. Classification technologies based on text analytics can automatically identify studies relevant to a given AOP from a very large corpus of literature. By greatly reducing the time and effort required to screen documents, text analytics can increase the viability of large-scale efforts to assess the validity of a hypothesized AOP.

Generic Schematic of an Adverse Outcome Pathway

Toxicant	Macro-Molecular Interactions	Cellular Responses	Organ Responses	Organism Responses	Population Responses
▪ Chemical properties	▪ Receptor/ ligand interaction ▪ DNA binding ▪ Protein oxidation	▪ Gene activation ▪ Protein production ▪ Altered signaling	▪ Altered physiology ▪ Disrupted homeostasis ▪ Altered tissue development/ function	▪ Lethality ▪ Impaired development ▪ Impaired reproduction ▪ Cancer	▪ Structure ▪ Recruitment

Assessing the Potential of Text Analytics in AOP Development

Scope and Purpose of Simulation

- To assess the potential of text analytics in gathering supporting evidence for hypothesized AOPs, we worked backwards from a well-known published AOP relating to liver fibrosis that is available on the AOP Wiki (see figure below).

Approach

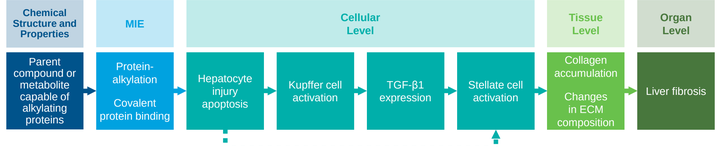
- Scraped the actual AOP-supportive citations from the published AOP as a measure of retrieval efficiency using automated technology.
  - The citations in the AOP PDF were inconsistently formatted, so we used the open-source Freecite tool to parse titles. Although this process was automated, the lack of formatting of references in the original document resulted in some level of noise (imperfectly parsed titles).
  - Using these citation titles as a search term in the PubMed API, we pulled the associated PMIDs and abstracts. This process was also imperfect and resulted in a slightly greater number of references returned than in the original AOP.
- Gathered a dataset of approximately 4 million studies (PMIDs, titles, and abstracts) from the PubMed API related to "mode of action" and "adverse outcome pathway."
- Used keywords from the AOP summary table and applied several text analytics tools and approaches to identify studies potentially relevant to the AOP.

Validation Approach

- Compared the PMIDs that we retrieved from the PubMed API against the PMIDs in the AOP citation list.
- Experts reviewed and assessed the relevance of a sample of the retrieved abstracts that we retrieved from the PubMed API and that were not in the original AOP citation list.

AOP Test Case: Protein Alkylation Leading to Liver Fibrosis

Landesmann, B. et al. published on AOP Wiki (https://aopwiki.org/aops/38)



1 Customized Keyword Search

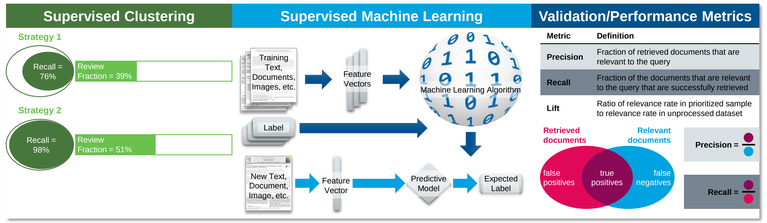
- **Approach:** Using the keywords in the AOP Table (see right) we conducted advanced keyword searches on a corpus of nearly 4 million abstracts. Multiple paired keyword searches were performed for each stepwise link in the AOP.
- **Proximity Search:** We assumed that hypothesized causative links between points in the AOP could be validated by finding literature in which the associated AOP terms occurred within a limited proximity of one another. Pairwise proximity searches for terms defining successive links in the AOP were performed using the DoCTER tool.
- **Fuzzy Search:** To account for the possibility that documents may refer to slightly varying grammatical forms of the same term, such as alkylation and alkylating, we enabled the use of fuzzy string searches in the DoCTER tool. Fuzzy searches return character strings that are approximately equal to the search term using a metric known as the Levenstein distance.
- **Retrieval Strategies:** For our default retrieval strategy we used a proximity of 200 characters (roughly within a 20 word distance) and Levenstein distance of 2 for each AOP term-pair. This resulted in a precise dataset of only 322 abstracts. Less rigorous criteria yielded larger retrievals.

Keywords in Liver Fibrosis AOP

Title	Directness	Evidence
Alkylation, Protein leads to N/A, Cell injury/death	Directly leads to	Moderate
N/A, Cell injury/death leads to Increased, Activation and Recruitment of Hepatic macrophages (Kupffer Cells)	Directly leads to	Strong
N/A, Cell injury/death leads to Activation, Stellate cells	Indirectly leads to	Strong
Increased, Activation and Recruitment of Hepatic macrophages (Kupffer Cells) leads to Up Regulation, TGFbeta1 expression	Directly leads to	Strong
Up Regulation, TGFbeta1 expression leads to Activation, Stellate cells	Directly leads to	Strong
Activation, Stellate cells leads to Accumulation, Collagen	Directly leads to	Strong
Accumulation, Collagen leads to N/A, Liver fibrosis	Directly leads to	Strong

2 Supervised Clustering and Machine Learning

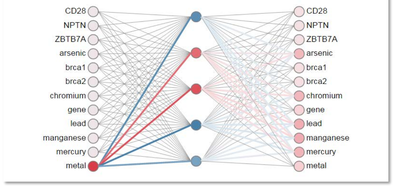
- **Approach:** To identify abstracts relevant to the AOP, we used a training (seed) dataset of 157 abstracts, developed partially by using the results of the customized keyword search strategy. ICF's DoCTER tool was deployed to implement both supervised clustering and machine learning.
- **Supervised Clustering:** By including positive and negative seeds with a larger set of 500,000 abstracts, we were able to identify subsets of abstracts meeting alternative precision and recall metrics. We eliminated approximately 50% of the corpus in this way.
- **Supervised Machine Learning:** To further prioritize abstracts identified by supervised clustering, we used the training dataset to rank the shortlisted abstracts in terms of relevance probability scores using machine learning algorithms.
- **Validation:** A random sample of the top decile of the prioritized abstracts were selected for expert review for relevance to the AOP.



3 Smart Clustering

- **Context:** Instead of letting clusters form where they will as in standard clustering, smart clustering forms a single ranked cluster around words or phrases known to be relevant to the search objective.
- **Semantic Similarity Algorithms:** Smart clustering uses semantic similarity algorithms to help users develop a set of topic defining keywords. By representing words in a corpus of text as mathematical vectors, semantic similarity algorithms are able to generate synonymous words and phrases and generate association scores. ICF's ToxSyn tool deploys these algorithms to enable the generation of synonyms for purposes such as keyword expansion in literature search.
- **Approach:** We used pairwise iterations of keywords defining each link in the AOP chain to generate a ranked set of abstract relevant to the AOP. We expanded this list of AOP-specific keywords by using ToxSyn. We did not manually curate the suggested list of related or synonymous terms proposed by ToxSyn. Nor did we differentially weight the alternative keywords, although these step are desirable.
- **Validation:** To assess the performance of smart clustering, we added our seed dataset to a dataset 500,000 abstracts drawn from the data discussed earlier. We estimated the lift metric for the top centile of ranked abstracts.

Generic Representation of Semantic Similarity Algorithms



4 Results

Strategy	Metric	Result
Proximity & Fuzzy Search	Retrieval Efficiency	Retrieved 57/199 of AOP citations list*
Proximity & Fuzzy Search	Precision	Precision = 10% for 265 abstracts not in AOP citations
Supervised Clustering	Decision Curve	49% Elimination for 98% Recall
Supervised Clustering and ML	Lift in Top Decile	Lift = 100 **
Smart Clustering	Lift in Top Centile	Average Lift = 32 **

*Actual value may be higher
** Response rate in general population may be subject to sampling error

Proximity and Fuzzy Search

Reduced 4 million abstracts to 322 abstracts, of which:

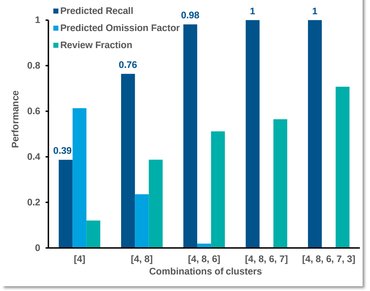
Supervised Clustering and Machine Learning

- Eliminates 49% with 2% relevance loss
- Combined with machine learning, prioritization results in top decile with a lift of approximately 100
- Can create concentrated cluster of high relevance abstracts

Smart Clustering

- Low cost method (no investment in training or expert review)
- Top centile shows average lift of approximately 30
- Better results for some combinations of AOP keywords than others
- Results likely to improve with manual vetting of automatically generated keywords and weighting.

Decision Curve based on Supervised Clustering



Conclusions, Caveats, and Future Directions

- The text analytics methods used in this simulation successfully identified a high percentage of literature relevant to the hypothesized AOP with relatively low effort.
- High levels of performance – measured by recall, precision and lift -- were achieved in the context of the simulation conditions.
- A relatively small training/seed dataset was used in the analysis. Predictions and performance are likely to be enhanced with a larger training dataset.
- A vast starting set of 4 million abstracts was used here to demonstrate computational power. Strategic reduction of the initial dataset would greatly enhance performance.
- In future, these approaches could be automated to support large-scale evidence gathering and literature prioritization for computationally hypothesized AOPs.
- Use of full text articles instead of abstracts is likely to increase the accuracy of document retrieval

DOCTER
Document Classification and Topic Extraction Resource

ToxSyn
Toxicological Semantic Similarity Discovery

Logo Design and Graphics: ***Courtney Skuce***