Unsupervised Identification of Study Descriptors in Toxicology Research

[1]**Dasha Herrmannova (herrmannovad@ornl.gov)**, [1]Steven R. Young, [1]Robert M. Patton,
[1]**Christopher G. Stahl (stahlcg@ornl.gov)**, [2]Nicole C. Kleinstreuer, [2]Mary S. Wolfe

[1]Oak Ridge National Laboratory, TN, USA
[2]NIEHS, NIH, Research Triangle Park, NC, USA

1. Motivation

- Extracting data elements from publication full texts is an **essential step in a number of tasks**, however, at present, it is **time-consuming, largely manual and requires domain expertise**
- Typical approach to automated data extraction is to build a prediction model from training data – however, there are two issues with this approach:
  - Obtaining training data can be very costly
  - Depending on task, data being extracted can vary significantly
- Therefore, we focus on **unsupervised methods for identifying text segments relevant to the information being extracted**

The intact female weanling version in the Organization for Economic Cooperation and Development (OECD) uterotrophic assay Test Guideline (TG) 440 is proposed as an alternative to the adult ovariectomized female version, because it does not involve surgical intervention (vs the ovariectomized version) and detects direct/indirect-acting estrogenic/anti-estrogenic substances (vs the ovariectomized version which detects only direct-acting estrogenic/antiestrogenic substances binding to the estrogen receptor). This validation study followed OECD TG 440, with six female weanling rats (postnatal day 21) per dose group and six treatment groups. Females were weighed and dosed once daily by oral gavage admin. route for three consecutive days, with one of six doses of 17α-ethinyl estradiol in corn oil at 5 ml kg^{-1} at 0 and 0.1–10 μg kg^{-1} per day . On postnatal day 24, the juvenile females were euthanized by CO_2 asphyxiation, weighed, livers weighed and uteri weighed wet and blotted. The presence or absence of vaginal patency was recorded. Absolute and relative (to terminal body weight) uterine wet and blotted weights and uterine luminal fluid weights were significantly increased at 3.0 and 10.0 (both P $<$ 0.01) μg kg^{-1} per day, and increased to $\sim$140% of control values at 1.0 μg kg^{-1} per day (not statistically significantly). In vivo body weights, weight changes, feed consumption, liver weights and terminal body weights were unaffected. Vaginal patency was not acquired in any female at any dose, although vaginal puckering was observed in one female at 10.0 μg kg^{-1} per day. Therefore, this intact weanling uterotrophic assay is validated in our laboratory for use under US and European endocrine toxicity testing programs/legislation.

GS; GS; group size (GS); animal model; no. of dose groups; dosing interval; dose groups (neg. and pos. control); necropsy timing

Figure 1: Manually annotated text excerpt from a research publication taken from [1].

2. Data

- A reference **database of rodent uterotrophic bioassay data extracted from 670 research publications** [1]
  - The database was created to facilitate the development of novel *in vitro* methods for testing chemicals
- The studies in the database were assessed according to their adherence to test guidelines set forth in [2]:
  - Defined using six minimum criteria (MC) (Table 1)
  - Guideline-like studies (GL)**: all six MC have to be met**
- Preparing the database **took two people two years**, therefore, **significant time and resources could be saved by automating the process**

Table 1: Minimum criteria for guideline-like studies (shortened). Source: [1].

Minimum criteria	Description
1: Animal model	Immature rats, ovariectomized (OVX) adult rats, or OVX adult mice are acceptable (immature mice are not acceptable). OVX animals: OVX should be performed between ...
2: Group size	Each control group should have a minimum of three animals and each test group should have a minimum of five animals.
3: Route of administration	Acceptable routes of administration: oral gavage (p.o.), subcutaneous (s.c.) injection, or intraperitoneal (i.p.) injection.
4: Number of dose groups	Minimum of two dose level groups. Must have positive control and negative control.
5: Dosing interval	Dosing for a minimum of three consecutive days. Complete by PND 25 in immature animals.
6: Necropsy timing	Should be carried out 18-36 hours after the last dose.

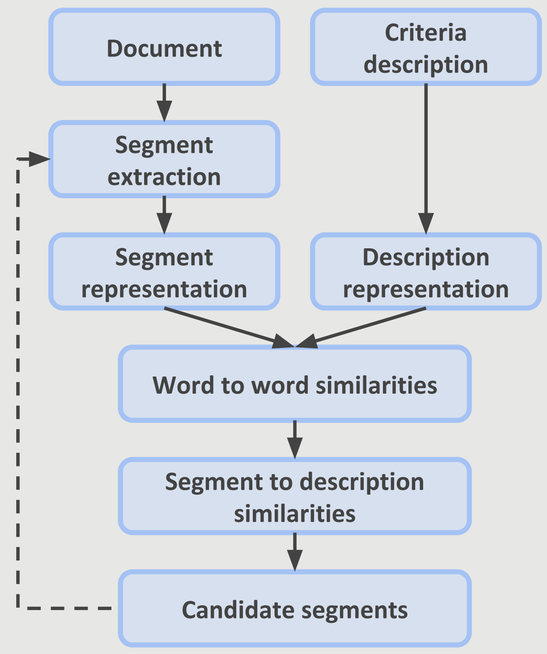
- The database contains **670 publications** (but **only 97 (~14%)** contain GL studies)
  - Each study is assigned a 0/1 label for each MC (0=MC not met, 1=MC met)
  - There exist **no fine-grained text annotations showing where in text were the criteria mentioned**
- Most publications **contain multiple studies**
  - We don't distinguish between publications describing single or multiple studies
  - For each MC, if a document contained multiple studies with different labels (both 0s and 1s), we discarded that document from our analysis of that criteria
- Table 2 shows final label statistics for each of the criteria (after cleanup)

Table 2: Label statistics.

Criteria	Criteria not met	Criteria met	Total	% of positive
MC 1	414	175	589	29.71
MC 2	35	577	612	94.28
MC 3	70	536	606	88.45
MC 4	309	206	515	40.00
MC 5	96	490	586	83.62
MC 6	228	340	568	59.86
GL	**522**	**72**	**594**	**12.12**

3. Approach

- The intuition is **based off question answering systems**
- We treat the **criteria descriptions** (Table 1) **as the question** and the **text segments** within the publication full text that discuss the criteria **as the answers**
- The goal is to **find the text segments most likely to contain the answer**
- We represent the criteria descriptions and text segments extracted from the documents as vectors of features, and utilize relatedness measures to retrieve text segments most similar to the descriptions
- We represent words as vectors generated using Word2Vec
- A **high level overview** of our approach is shown in Figure 2



1. **Input:** Document and MC description
2. **Segment extraction:** Input document broken down into shorter sequences (e.g. sentences)
3. **Representation:** Text (segment from doc./description) represented as sequence of Word2Vec word vectors
4. **Word2Word similarities:** *Cosine similarity* between words from description and words from document segments
5. **Segment2Description similarities:**
   1. Select max similarity for each word in the document segment
   2. Segment similarity = average of the word similarities
6. **Candidate segments:** Top *k* most similar segments

Figure 2: High level overview of our approach

4. Evaluation

- Figure 3 shows an example annotation generated using our approach

The intact female weanling version in the Organization for Economic Cooperation and Development (OECD) uterotrophic assay Test Guideline (TG) 440 is proposed as an alternative to the adult ovariectomized female version, because it does not involve surgical intervention (vs the ovariectomized version) and detects direct/indirect-acting estrogenic/anti-estrogenic substances (vs the ovariectomized version which detects only direct-acting estrogenic/anti-estrogenic substances binding to the estrogen receptor). This validation study followed OECD TG 440, with six female weanling rats (postnatal day 21) per dose group and six treatment groups. Females were weighed and dosed once daily by oral gavage for three consecutive days, with one of six doses of 17alpha-ethinyl estradiol in corn oil at 5 ml kg(-)(1) at 0 and 0.1-10 microg kg(-)(1) per day. On postnatal day 24, the juvenile females were euthanized by CO(2) asphyxiation, weighed, livers weighed and uteri weighed wet and blotted. The presence or absence of vaginal patency was recorded. Absolute and relative (to terminal body weight) uterine wet and blotted weights and uterine luminal fluid weights were significantly increased at 3.0 and 10.0 (both P < 0.01) microg kg(-)(1) per day, and increased to ~140% of control values at 1.0 microg kg(-)(1) per day (not statistically significantly). In vivo body weights, weight changes, feed consumption, liver weights and terminal body weights were unaffected. Vaginal patency was not acquired in any female at any dose, although vaginal puckering was observed in one female at 10.0 microg kg(-)(1) per day. Therefore, this intact weanling uterotrophic assay is validated in our laboratory for use under US and European endocrine toxicity testing programs/legislation.

Figure 3: Annotations generated using our method for MC 1. Abstract is from Figure 1.

- Goal: explore whether our approach truly identifies mentions of the MC in text
- We have utilized the existing 0/1 labels to train one binary classifier for each MC
- We have then compared three models which utilized selected sentences:
  1. ***k* sentences most similar** to the given MC
  2. ***k* least similar sentences**
  3. ***k* randomly selected sentences** (but none of the top or bottom *k* sentences)
- Intuition: a classifier utilizing the correct sentences should perform better

Table 3: Evaluation results.

Approach	MC 1	MC 2	MC 3	MC 4	MC 5	MC 6
Top *k*	**76.84**	91.55	87.71	**68.35**	**88.54**	**74.23**
Random *k*	73.23	**93.72**	**88.43**	65.65	85.29	68.28
Bottom *k*	70.00	91.39	88.23	63.10	80.60	63.70

- For four of the six criteria the top *k* model performs best
- A possible explanation for the top *k* model not performing best in the case of MC 2 and MC 3 is class imbalance (the top *k* sentences may not contain enough negative examples to learn from)

5. References

1. Kleinstreuer et al. (2016). *A curated database of rodent uterotrophic bioactivity.* In Environmental Health Perspectives, 124(5)
2. OECD. (2007). *Test No. 440: Uterotrophic Bioassay in Rodents.* In OECD Guidelines for the Testing of Chemicals, Section 4.

Support for this research was provided by a grant from the National Institute of Environmental Health Sciences (AES 16002-001), National Institutes of Health to Oak Ridge National Laboratory. This research was supported in part by an appointment to the Oak Ridge National Laboratory ASTRO Program, sponsored by the U.S. Department of Energy and administered by the Oak Ridge Institute for Science and Education. This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).