Cochrane Medical Database Based Prediction

Battushig Myanganbayar
Supervisors: Prof. Regina Barzilay, Franck Dernoncourt

Massachusetts Instituteof Technology
[regina, franck]@csail.mit.edu, btushig@mit.edu

Abstract

In medical field, It is often the case that there are many research papers related to certain hypothesis, and it is hard for doctors to evaluate everyone of them to reach conclusive decision about the hypothesis.

Cochrane Medical Database

- Collective summary of *6,000 medical hypothesis*
- *Most credible*, and widely used database by doctors
- *Expensive*, and *slow* review process
- Vast number of hypothesis has not been evaluated yet

1. Introduction / Structure of Data

Cochrane Article

Each Cochrane article has following subsection (example provided from one article)

- OBJECTIVES:

To evaluate the impact of **cancer genetic risk-assessment services** on patients at **risk of familial breast cancer**.

- MAIN RESULTS:

In this review update, we included five new trials, bringing the total number of included studies to eight. The included trials (pertaining to 10 papers), provided data on 1973 participants and assessed the impact of cancer genetic risk assessment on outcomes including perceived risk of inherited cancer, and psychological distress. **This review suggests that cancer genetic risk-assessment services help to reduce distress, improve the accuracy of the perceived risk of breast cancer, and increase knowledge about breast cancer and genetics**.

- AUTHORS' CONCLUSION:

This review found favourable outcomes for patients after risk assessment for familial breast cancer. However, there were too few papers to make any significant conclusions about how best to deliver cancer genetic risk-assessment services. Further research is needed assessing the best means of delivering cancer risk assessment

Referenced Articles

Each Cochrane article has at least **5 referenced** papers from following sources

- PubMed (because of easiness to scrape used for this study)
- CrossRef (Not used because of PDF Parsing)

Each referenced article on PubMed has following subsections:

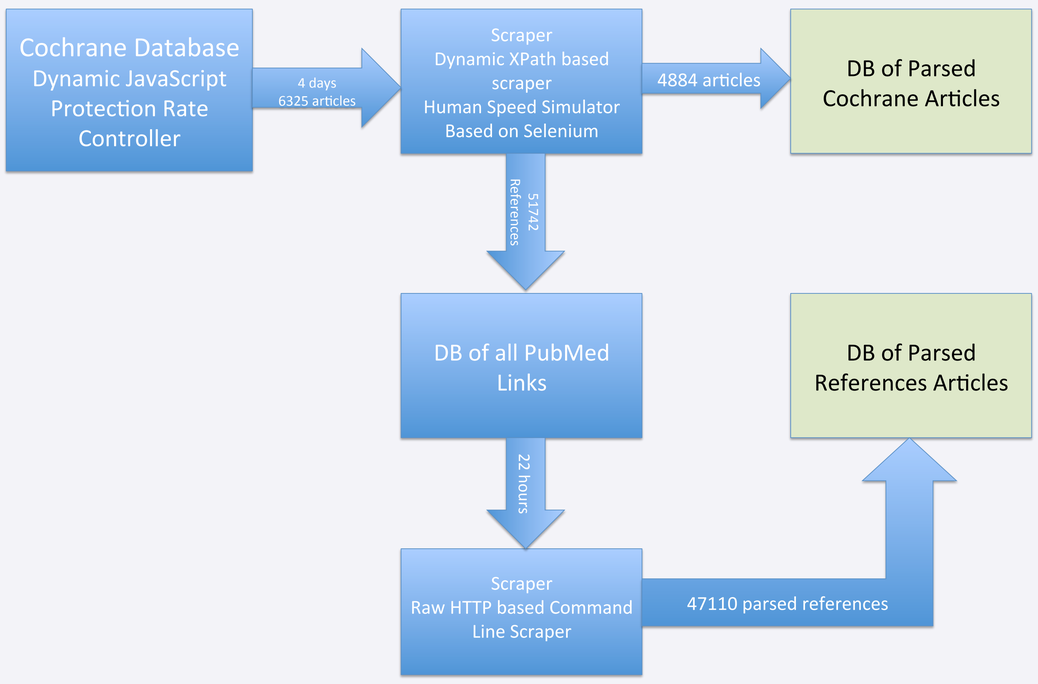
- RESULTS:

Although statistically significantly greater improvement in knowledge about breast cancer was found in the trial group (P: =.05), differences between groups in other psychologic outcomes were not statistically significant. **Women in both groups experienced statistically significant reductions in anxiety and found attending the clinics to be highly satisfying.** An initial specialist genetic assessment cost pound 14.27 (U.S. $22.55) more than a consultation with a breast surgeon.

- CONCLUSIONS:

There may be benefit in providing specialist genetics services to all women with a family history of breast cancer. Further investigation of factors that may mediate the impact of genetic assessment is in progress and may reveal subgroups of women who would benefit from specialist genetics services.

2. Data Set / Detailed View

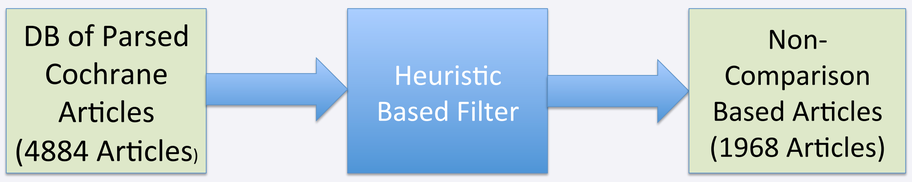


For some Cochrane article, there was insufficient amount of PubMed articles downloaded. Therefore, only Cochrane Articles with more than 5 referenced PubMed article is considered for the future section.

This criterion significantly reduced the number of eligible articles going from 4884 downloaded to only 1390 articles.

Each Cochrane Article is also labeled with ***"YES", and "NO"*** meaning either it supports the objective or not.

3. Algorithm Design



Objectives:

- To evaluate the effect of adjuvant RT following RP for prostate cancer in men with high risk features **compared with RP**. (Bad)
- To evaluate existing knowledge of the effect of antioxidants such as vitamin C, … fibrosis lung disease.
(Good objective)

Removes the comparison objected articles by looking at the phrases like "compare", "evaluate", "access", and their relative ordering

Objectives:

- To evaluate existing knowledge of the effect of antioxidants such as vitamin C, vitamin E, f-carotene, selenium and glutathione in cystic fibrosis lung disease.
(Good objective)

Feature Extraction from Reference Article:

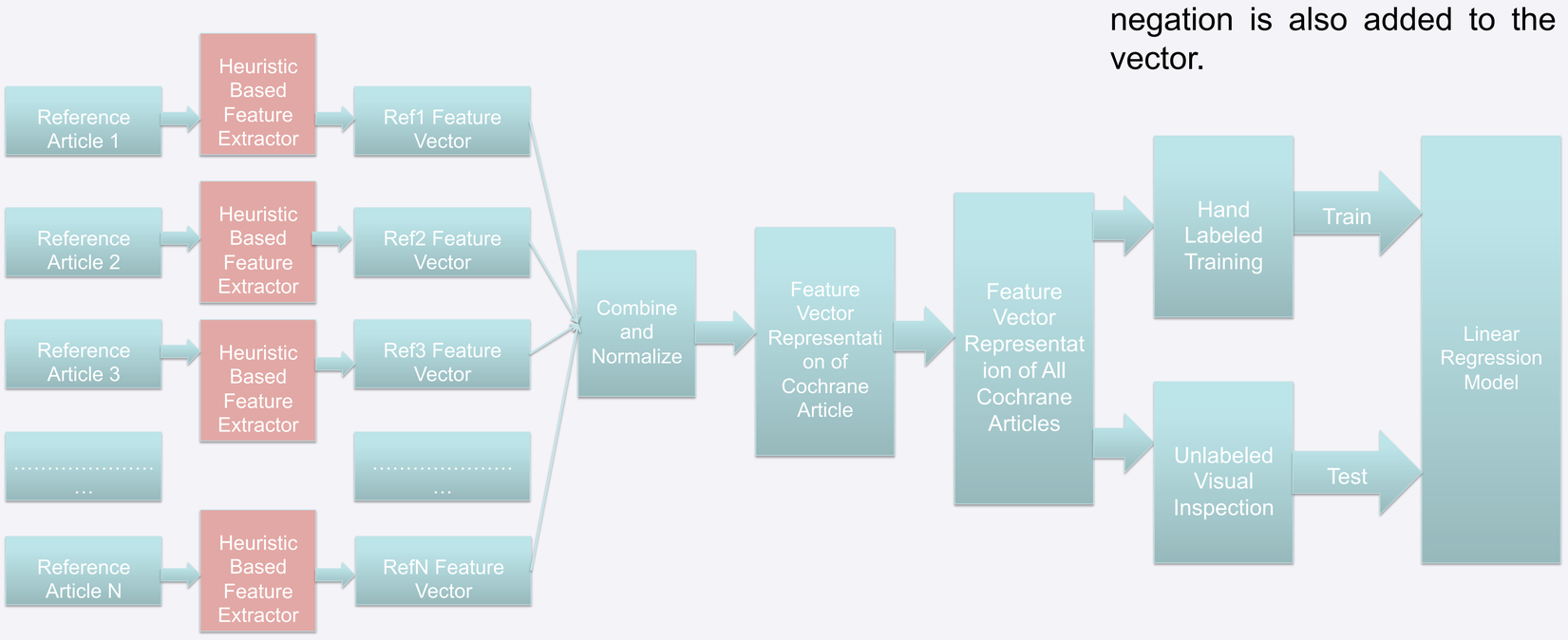
Heuristic Based Feature Extractor

To represent most of the information about reference article as being no and yes, following words are served as a heuristics, and one hot vector of 46 generated:

Action Verbs: support, improve, indicate, effect, affect, remove, reduce, decrease, benefit, less, increase

Adjectives: significant, potential, "well", "statistical", "strong", "substantial", "difference", "evidence", "better", "unclear", "safe"

For each 23 heuristics word it's negation is also added to the vector.



5. Results

We got 59.03% accuracy with 10 folds of cross validation for the dataset with 129 articles on the training set. Main comparison metric is based on random since only binary decision are being made. Distribution of Y/N is 49.71/50.29

	Acc	TP	FP	TN	FN
Random	0.497	0.242	**0.254**	0.257	**0.0018**
LR	**0.590**	**0.281**	0.214	**0.309**	0.0007

6. Further Improvements

More supervised data is crucial for the success of this project since meaning hidden inside the medical articles can vary a lot. We are also planning to try following new approaches in terms of classification:

- **Word Embedding** based feature vector generator instead of using heuristics
- ***Semantics of PubMed References*** instead of using unsupervised feature vector of PubMed articles, we can take majority vote from each reference article based on it's semantics.