Analyzing Inclusion Criteria of 7000 Cochrane Systematic Reviews

Presenters: Jingyi Xie¹& Xiaoru Dong¹; Advisor: Prof. Jodi Schneider²; Mentor: Linh Hoang² 1 - Department of Statistics, College of Liberal Arts and Sciences, University of Illinois at Urbana-Champaign; 2 – School of Information Sciences, University of Illinois at Urbana-Champaign

INTRODUCTION

We are interested in building a machine learning classifier to distinguish sentences like these:

"Only randomized trials were included..." ("Only RCTs")

"Randomized and quasi-randomized controlled trials were included..." ("Others")

DEFINITIONS

Systematic review: A particular kind of literature review.

- review, such as types of a trial. **Randomized controlled trial (RCT):** The people
- receiving treatment and a control group.
- Weka: A machine learning toolkit. Weka uses different algorithms to train a classifier based on our annotated data.

METHOD

Annotate data manually: ground truth	 1756 lines till nov First stage: two p the agreement
Propose features and implement codes to extract features	 Write Python co The frequency of feature
Use Weka to run classification	 Cross validatio Algorithms "Ra Bayes Classifie
Evaluate the results by comparing to the	 Values of Prec three algorithm

Inclusion criteria: Characteristics that the prospective trials must have if they are to be included in the systematic

participating in the trial are randomly allocated to a group

Cochrane: The group conducts systematic reviews of health care interventions publishes them in the Cochrane Library.



"Only R Others Weighte

Results from: Weka (RandomForest)

"Only R Others Weighte

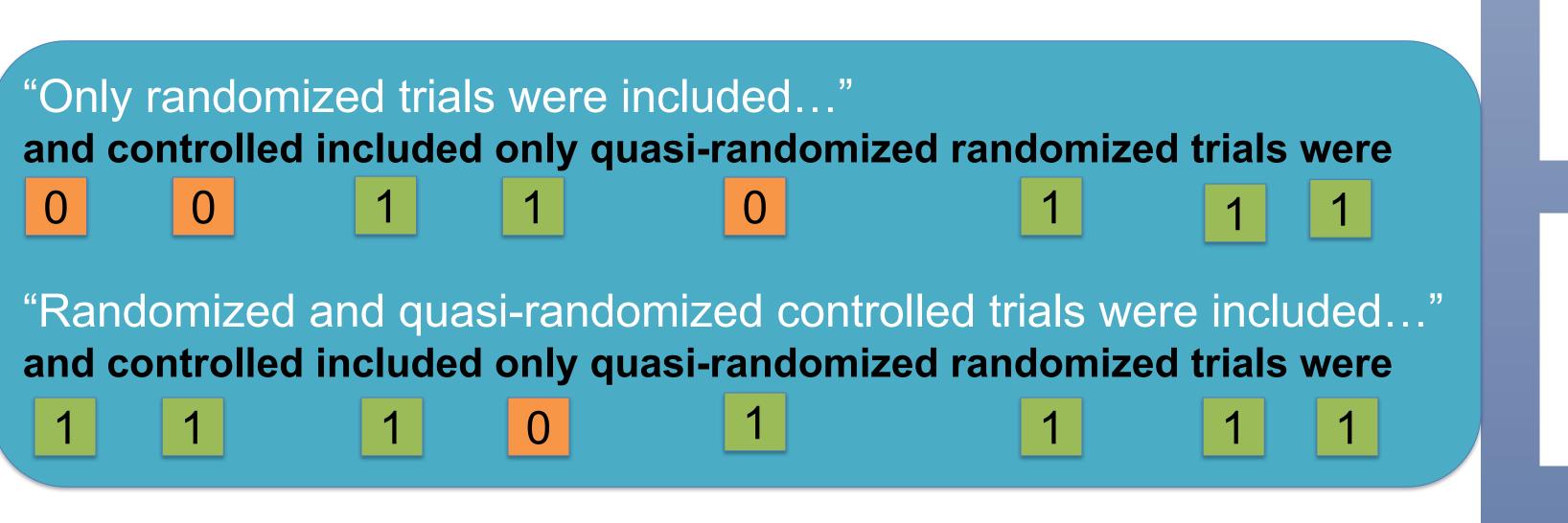
"Only RC Others Weighte

ow have been analyzed (by two people) people annotate same data to check More than 90% of the annotated results are the same!

ode to count the words of some word can be a

on: 10 times RandomForest" "Naïve ier" and "J48" are used

cision and Recall from the ms are compared



Seen from the "Result"

	Precision	Recall
CT"	81.8%	96.3%
	90.7%	62.4%
ed Avg.	85.0%	84.0%

	Precision	Recall
CT"	80.9%	78.7%
	64.4%	67.3%
ed Avg.	74.9%	74.6%

Results from: Weka (Naive Bayes Classifier)

	Precision	Recall
CT"	87.4%	91.1%
	83.2%	77.1%
<mark>ed Avg.</mark>	85.9%	86.0%

Results from: Weka (J48)



FUTURE WORK

- truth.

ACKNOWLEDGEMENTS

Linh Hoang is funded by National Library of Medicine: "Text Mining Pipeline to Accelerate Systematic Reviews in Evidencebased Medicine" (R01LM010817).

Thank you to Cochrane for providing Cochrane reviews as machine readable XML from which inclusion criteria were extracted.

Weka





School of Information Sciences

 Finishing annotations for the whole data set (7500 reviews).

 Right now each annotator works in different portions of the data set. We want to have two annotators annotate the whole data set independently and determine internal annotation agreement after that. This would help us to access the quality of the ground

 Doing in-depth analysis to understand which algorithm works best for the data and why.

 Analyzing and doing features selection (e.g. which words are informative for the classification).

REFERENCES

https://www.cs.waikato.ac.nz/ml/weka/

