Comparative Analysis of Association Rule Mining, Crowdsourcing, and NDF-RT Knowledge Bases for Problem-Medication Pair Generation

Karthik Sethuraman,[1] Allison B. McCoy, PhD[2], Dean F. Sittig, PhD[2]

1. Department of Bioengineering, School of Engineering and Applied Sciences, University of Pennsylvania
2. National Center for Cognitive Decision Making in Healthcare
and School of Biomedical Informatics, University of Texas Health Science Center at Houston

UTHealth The University of Texas Health Science Center at Houston School of Biomedical Informatics

SHARPC NCCD

CPRIT CANCER PREVENTION & RESEARCH INSTITUTE OF TEXAS

Objective

Generally, to better compile and organize the growing patient information confronting healthcare providers through automatic summarization.

Specifically, to evaluate three different approaches, association rule mining, crowdsourcing, and the National Drug File-Reference Terminology (NDF-RT), to problem-medication pair generation – a key auto-summarization task for clinical datasets.

Data and Knowledge Bases

Data: Electronic health record (EHR) data from a large, multi-specialty, ambulatory academic practice that provides medical care for all ages in Houston. One year study period from June 1, 2010 to May 30, 2011, including 418,221 medications and 1,222,308 problems logged for 53,108 patients.

Knowledge Base Generation: We independently applied association rule mining and crowdsourcing to the dataset to generate problem-medication knowledge bases (KBs) and mapped local EHR terminology to NDF-RT terminology for 3rd KB.

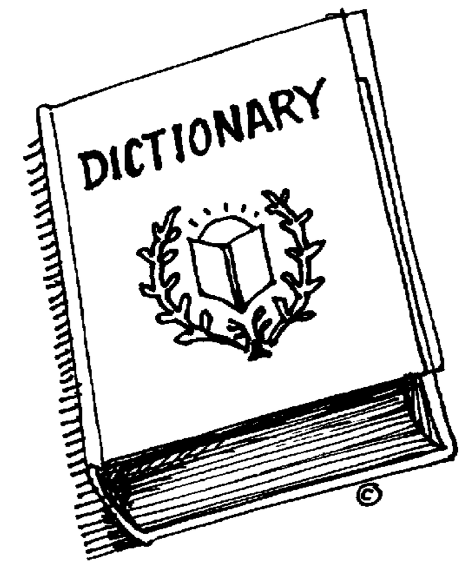
Analysis

We sorted each KB by appropriateness measures and examined the top 5,000 pairs per KB to eliminate spurious problem relations. We assessed similarities and differences between three KBs by:

i) constructing contingency tables by KB with presence/absence in two remaining KBs as margin criteria to determine degree of overlap (Fisher's test)

ii) measuring Spearman's rank correlation of overlapping problem-medication pairs across KBs to check if KBs ranks pairs in the same way indicating similar KB construction

iii) individual assessing the top fifty problem-medication pairs per KB



Association Rule Mining

Data mining technique that looks for relationships as co-occurrence of pairs in a database, here logged medication-problem information

Crowdsourcing

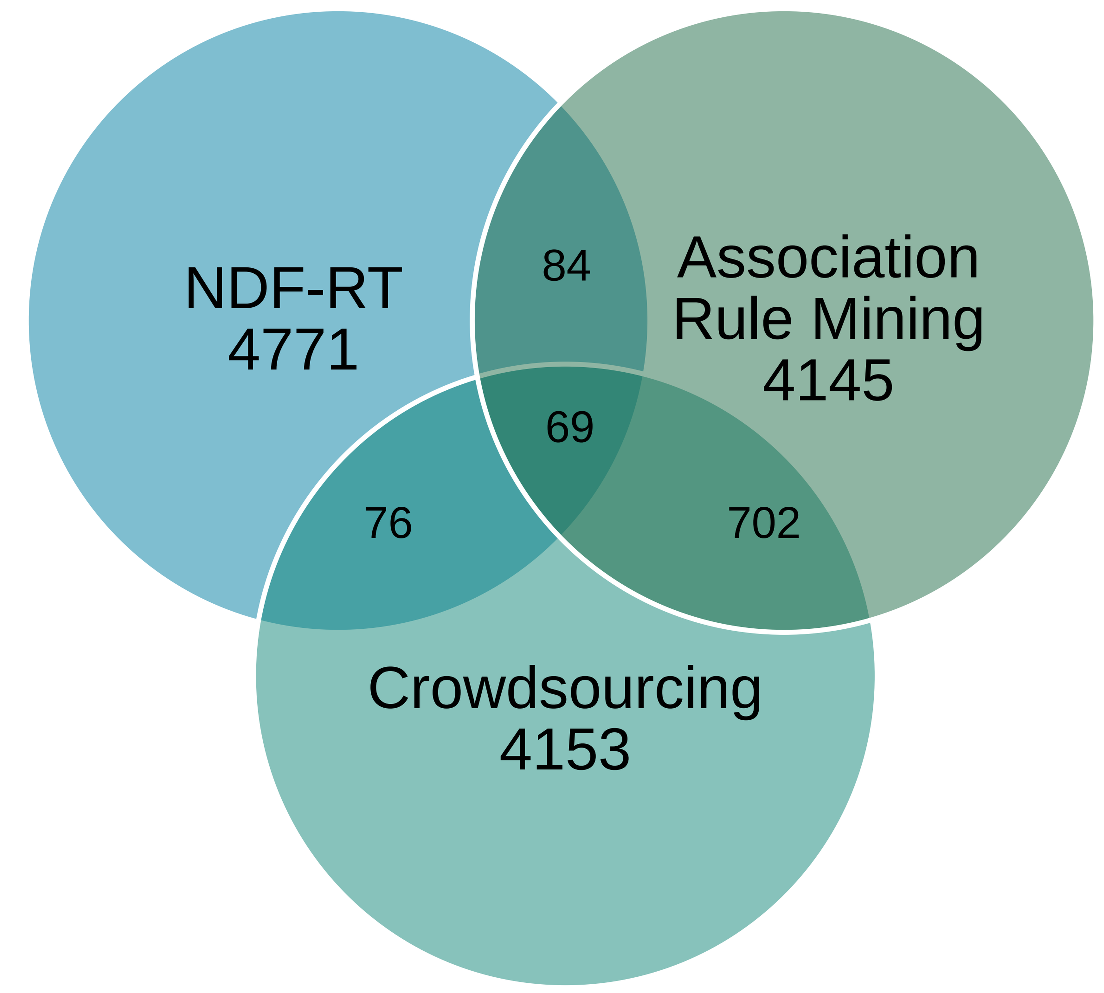
Using information gathered from a group or crowd, here the medications and problems entered by a set of identified physicians

NDF-RT

Curated reference of medications and associated information, including related problems/diseases.

Overlapping Pairs between KBs

Association Rule Mining and crowdsourcing have more overlapping pairs (771) than either with the NDF-RT (145, 153).



	Association Rule Mining		Crowdsourcing		NDF-RT	
Rank	**Problem**	**Medication**	**Problem**	**Medication**	**Problem**	**Medication**
1	Scabies	Permethrin	Hypo-thyroidism	Levothyroxine Sodium	Hypertension	Cardizem SR
2	Bacterial Vaginosis	Metronidazole	Hyperlipid-emia	Simvastatin	Hypertension	Cardene
3	Motor Neuron Disease	Rilutek	Hypertension	Lisinopril	Hypertension	Reserpine
4	Vaginal Candidiasis	Terconazole	Bacterial Vaginosis	Metronidazole	Hypertension	Triamterene
5	Pseudo-monas	Amikacin Sulfate	Hyperlipidimia	Lipitor	Hypertension	Innopran XL
6	Type 1 Diabetes	Glucagon	Hypertension	Hydrochloro-thiazide	Hypertension	Bendroflu-methiazide
7	Hypothyroid-ism	Levothyroxine Sodium	Hypertension	Amlodipine Besylate	Hypertension	Isradipine
8	Mitochondrial Metabolism Disorder	Levocarnitine	Allergic Rhinitis	Fluticasone Propionate	Hypertension	Eplerenone
9	Tinea Capitis	Griseofulvin (Microsize)	Esophageal Reflux	Nexium	Hypertension	L-Tyrosine
10	Congenital Adrenal Hyperplasia	Solu-Cortef	Type 1 Diabetes	Metformin HCl	Hypertension	Diltiazem HCl CR

Top Ten Problem-Medication Pairs

Association rule mining forms very sensitive but low frequency pairs. It finds very specific but more rare relationships like Scabies to Permethrin.

NDF-RT gives high-frequency pairs – all problems listed are hypertension.

Crowdsourcing provides a mixture of association rule mining and NDF-RT relations. 3/10 problems are hypertension but more rare relations like Allergic Rhinitis and Fluticasone Propionate are also present.

Fisher's Test and Spearman's σ

Fisher's test:
Overlap between association rule mining and crowdsourcing was very significant ($p<0.001$) while overlap between either of the two and NDF-RT was not ($p>0.10$).

Spearman's σ:
Association rule mining and crowdsourcing overlapping pairs displayed significant monotonic correlation (for ranks) with σ of 0.315 ($p<0.001$), but neither displayed correlation with NDF-RT.

Conclusions

Association rule mining and crowdsourcing are remarkably similar in pair generation.

The NDF-RT doesn't overlap with either.

Significant Spearman's σ and pair overlap between association rule mining and crowdsourcing indicate similar operation.

Similar relation formation may be because both depend on the local EHR dataset.

NDF-RT presents a distinct, expert-generated, non-dataset relation source that can be exploited for automatic summarization.

Generally, expert-generated (or reference) information may provide an underexploited source of relations for auto-summarization.

Acknowledgements

This project was supported by Grant No. 10510592 for Patient-Centered Cognitive Support under the Strategic Health IT Advanced Research Projects (SHARP) from the Office of the National Coordinator for Health Information Technology and by the CPRIT Summer Undergraduate Cancer Research Experience.

Please contact the corresponding author at: amccoy1@tulane.edu