

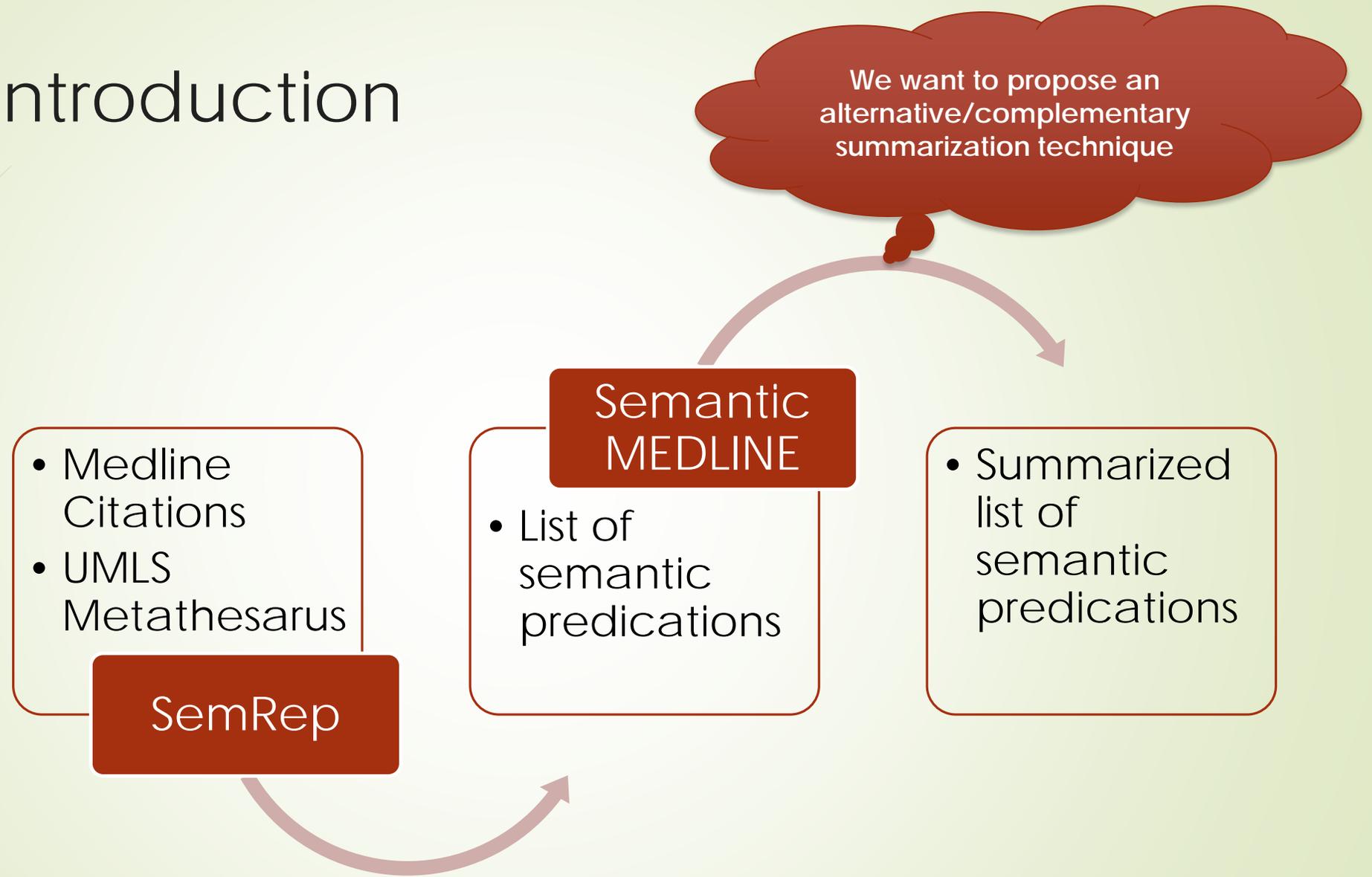
# A new approach for summarizing SemRep predications

Vahid Taslimitehrani

Dr. Olivier Bodenreider



# Introduction

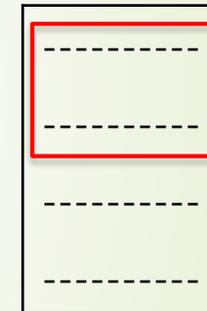


# Background

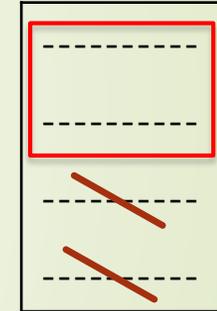
- ▶ Semantic MEDLINE summarization system
  - ▶ Semantic MEDLINE (2008)
    - ▶ Relevance
    - ▶ Connectivity
    - ▶ Novelty
    - ▶ Saliency
  - ▶ Degree Centrality (2011)
  - ▶ Clustering Cliques (2013)



Semantic  
MEDLINE



Our  
technique



Semantic  
MEDLINE +Our  
technique

# Motivation – an example

## UMLS Metathesaurus

Selective beta-1 adrenoceptor stimulants

Inferred

## SemRep semantic predications

|      |                             |            |                             |
|------|-----------------------------|------------|-----------------------------|
| IS_A | Dobutamine                  | TREAT<br>S | Congestive<br>heart failure |
| IS_A | Dopexamine                  | TREAT<br>S | Congestive<br>heart failure |
| IS_A | Dopexamine<br>hydrochloride | TREAT<br>S | Congestive<br>heart failure |
| IS_A | Xamoterol                   | TREAT<br>S | Congestive<br>heart failure |

Summarize

Selective beta-1  
adrenoceptor stimulants  
TREATS  
Congestive heart failure

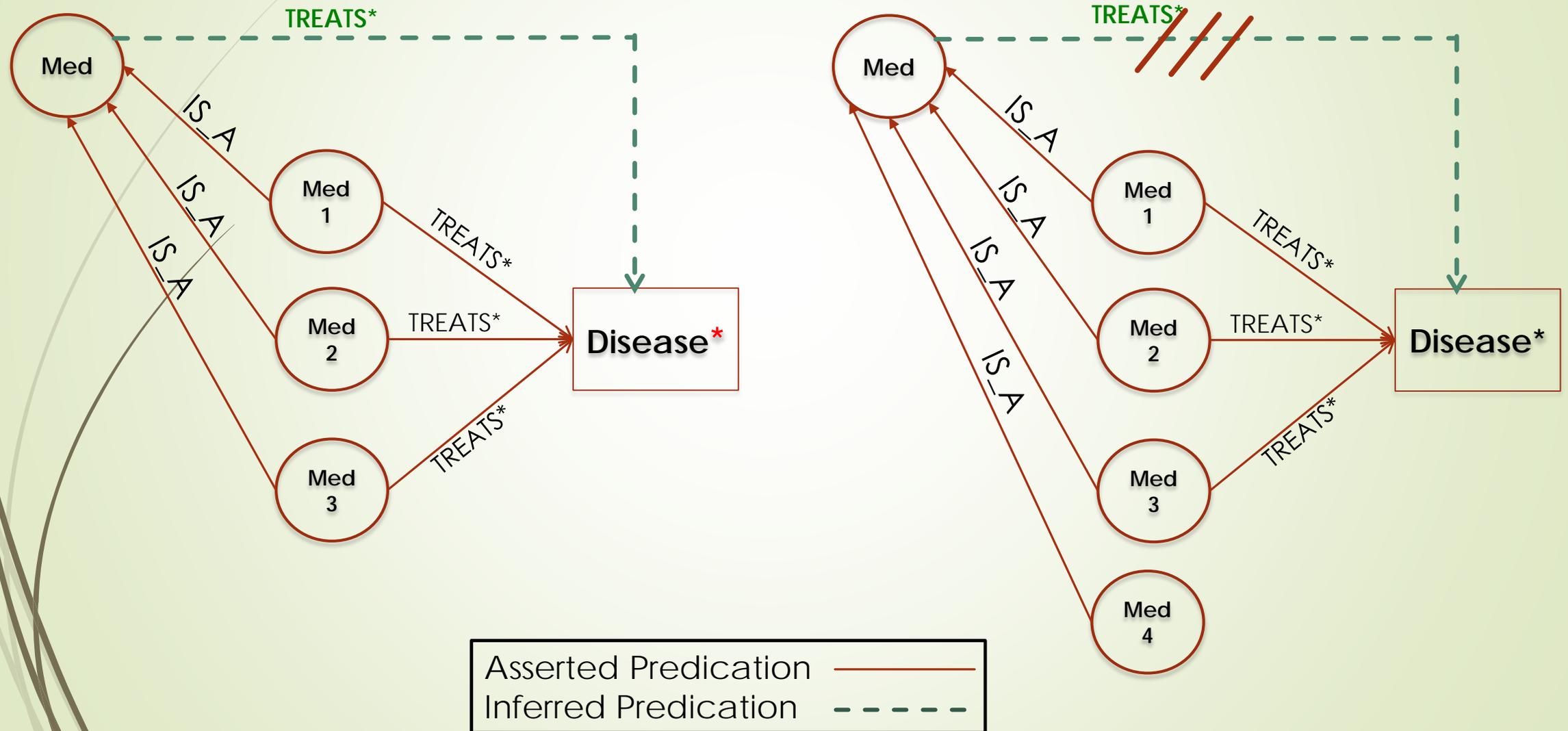
# Motivation-an example

- Based on the example, we observe:
  1. 4 semantic predications are aggregated.
  2. A **new inference** is made.

## Objective:

Our technique leverages hierarchical relations from the UMLS Metathesaurus for aggregating the semantic predications and generating new inferences from the aggregated predications.

# Methodology-an overview



# Methodology

- ▶ Let's discuss about the details of methodology using an example. We're interested to summarize the semantic predications returned in response to the following question:

What are the medications used to TREAT Congestive heart failure (C0018802)?

- ▶ SemRep returns 6013 semantic predications.
- ▶ SemRep returns 684 unique treatment options.

# Methodology

- ▶ **Step 1:** Retrieve unique semantic predications from SemRep when
  - ▶ The predicate is TREATS or any descendant.
  - ▶ The object is a disease or any descendant.

Congestive heart failure

- ▶ **Step 2:** Extract semantic groups of each subject and remove those predications with procedures as semantic group (T061, ...)
- ▶ **Step 3:** Retrieve all parents of each medication returned from step 2.

684 Semantic  
predications

500 Semantic  
Predications

**Xamoterol** is one of the medications from step 2 (**Xamoterol** TREATS **Congestive heart failure**)

Xamoterol has two parents:

1. Selective beta-1 adrenoceptor stimulants (**Xamoterol** IS\_A **Selective beta-1 adrenoceptor stimulants**)
2. Sympathomimetics (**Xamoterol** IS\_A **Sympathomimetics**)

# Methodology

9

- **Step 4:** Retrieve all children of the parents returned from step 3.

**Selective beta-1 adrenoceptor stimulants** has 4 children:

1. Dobutamine (Selective beta-1 adrenoceptor stimulants INVERSE\_ISA Dobutamine)
2. Dopexamine (Selective beta-1 adrenoceptor stimulants INVERSE\_ISA Dopexamine )
3. Dopexamine hydrochloride (Selective beta-1 adrenoceptor stimulants INVERSE\_ISA Dopexamine hydrochloride)
4. Xamoterol (Selective beta-1 adrenoceptor stimulants INVERSE\_ISA Xamoterol)

**Sympathomimetics** has many children such as:

1. Adrenergic alpha-agonists
2. Dopamine
3. Ephedrine
4. Etilefrine
5. Xamoterol

- **Step 5:** Some of the parents are too generic such as C0993159 (Oral product). If an ancestor has too many descendants, then it is not used.

# Methodology

- Step 6:** For each child of a parent returned from step 5, we need to verify the child TREATS the disease or not. If all children TREAT the disease, we aggregate semantic predications and make a new inference.

## Selective beta-1 adrenoceptor stimulants

|                                     |   |
|-------------------------------------|---|
| Dobutamine TREATS CHF               | ✓ |
| Dopexamine TREATS CHF               | ✓ |
| Dopexamine hydrochloride TREATS CHF | ✓ |
| Xamoterol TREATS CHF                | ✓ |

Aggregate

Selective beta-1 adrenoceptor stimulants  
TREATS  
 Congestive heart failure

## Sympathomimetics

|                       |   |
|-----------------------|---|
| Adrenergic TREATS CHF | ✓ |
| Dopamine TREATS CHF   | ✓ |
| Ephedrine TREATS CHF  | ✗ |
| Xamoterol TREATS CHF  | ✓ |

Do not aggregate

✗

# Methodology

- ▶ **Step 7:** If we can aggregate semantic predications in step 6, we continue to aggregate into the higher levels. If not, it is the highest level of summarization.

**Selective beta-1 adrenoceptor stimulants** has just one parent:

1. Adrenergic beta agonist

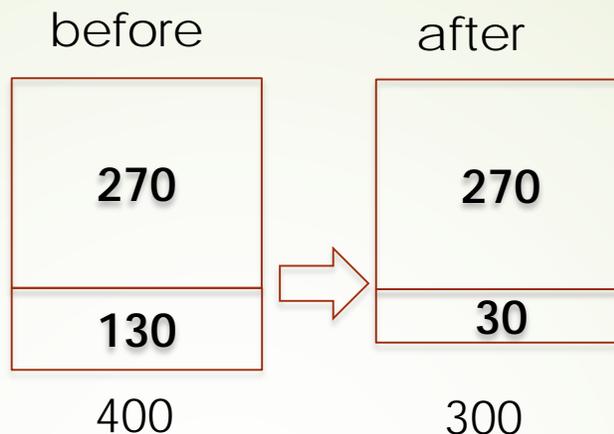
and **adrenergic beta agonist** has 6 children:

1. Selective beta-1 adrenoceptor stimulants ✓
2. Selective beta-2 adrenoceptor stimulants ✗
3. Dobutamine ✓
4. Dobutamine hydrochloride ✓
5. Isoproterenol ✗
6. Mirabegon ✗

# Implementation

- ▶ We used Biomedical Knowledge Repository (BKR) to implement our technique.
  - ▶ UMLS in RDF
  - ▶ SemRep predications in RDF
- ▶ Created at NLM by Dr. Olivier Bodenreider and Dr. Thomas Rindflesch.
- ▶ We used 2013 version that includes more than 27 million predications from 13 million articles by SemRep.
- ▶ Technologies
  - ▶ Semantic Web
    - ▶ RDF, SPARQL standards
    - ▶ Virtuoso triple store
  - ▶ Programming
    - ▶ Java

# Evaluation



- We defined 4 quantitative measures to evaluate the performance of our technique.

- Summarization rate:  $1 - \frac{\# \textit{semantic predications after}}{\# \textit{semantic predications before}} = 1 - \frac{300}{400} = 0.25$

- Inference ratio:  $\frac{\# \textit{predications can be aggregated}}{\# \textit{inferences}} = \frac{130}{30} \cong 4.3$

- Number of generated inferences
- Ratio of validated inferences

# Experimental results

- ▶ We investigate two questions:
  1. What are the medications used to treat disease X?
  2. What are the medications caused disease X? (adverse drug events)
- ▶ For each question, we select five diseases with
  - ▶ high numbers of predications (more than 400 unique predications)
  - ▶ medium numbers of predications (between 100 and 400 predications)

# Experimental results-question #1

| Disease                  | # pred. | Summarization rate | # generated inferences | Ratio of validated inferences | Inference ratio |
|--------------------------|---------|--------------------|------------------------|-------------------------------|-----------------|
| Hypertensive disorder    | 1122    | 26%                | 63                     | 38%                           | 5.6             |
| Congestive heart failure | 499     | 29%                | 24                     | 21%                           | 7               |
| Depression               | 400     | 27%                | 39                     | 23%                           | 3.7             |
| Myocardial infarction    | 419     | 29%                | 24                     | 16%                           | 6               |
| Schizophrenia            | 401     | 30%                | 26                     | 38%                           | 5.6             |

Diseases with high numbers of semantic predications (more than 400)

# Experimental results-question #1

| Disease                | # pred. | Summarization rate | # generated inferences | Ratio of validated inferences | Inference ratio |
|------------------------|---------|--------------------|------------------------|-------------------------------|-----------------|
| Hypercholesterolemia   | 240     | 21%                | 18                     | 33%                           | 3.8             |
| Pruitus                | 179     | 47%                | 12                     | 50%                           | 8               |
| Burn injury            | 316     | 45%                | 11                     | 55%                           | 13.9            |
| Pseudomonas Infections | 201     | 26%                | 23                     | 43%                           | 3.3             |
| Glaucoma               | 343     | 29%                | 21                     | 62%                           | 5.7             |

Diseases with the medium numbers of semantic predications

## Experimental results-question #2

| Disease                  | # pred. | Summarization rate | # generated inferences | Ratio of validated inferences | Inference ratio |
|--------------------------|---------|--------------------|------------------------|-------------------------------|-----------------|
| Traumatic Injury         | 1045    | 40%                | 83                     | 17%                           | 6               |
| Ischemia                 | 622     | 44%                | 32                     | 3%                            | 9.5             |
| Cerebrovascular accident | 401     | 34.5%              | 16                     | 6%                            | 9.6             |
| Obstruction              | 1815    | 37%                | 75                     | 16%                           | 10              |
| Septicemia               | 748     | 41%                | 42                     | 14%                           | 11.5            |

Diseases with high numbers of semantic predications (more than 400)

## Experimental results-question #2

| Disease                         | # pred. | Summarization rate | # generated inferences | Ratio of validated inferences | Inference ratio |
|---------------------------------|---------|--------------------|------------------------|-------------------------------|-----------------|
| Wounds & injuries               | 139     | 44%                | 6                      | 50%                           | 8.4             |
| Cardiovascular disease          | 170     | 37%                | 8                      | 25%                           | 6.7             |
| Pulmonary embolism              | 152     | 35%                | 7                      | 28%                           | 6.6             |
| Asthma                          | 192     | 41%                | 11                     | 36%                           | 11.3            |
| Gastroesophageal reflux disease | 101     | 32%                | 5                      | 60%                           | 14.8            |

Diseases with the medium numbers of semantic predications

# Limitations

- ▶ We need to validate the rest of new inferences by the experts.
- ▶ In the current experiments, we're using SNOMED\_CT and MEDCIN hierarchy and IS\_A relations are used to explore the hierarchy. The code has the ability to generalize to any other hierarchies.

# Conclusion

- ▶ We propose a new technique to summarize SemRep predications.
- ▶ Our technique is based on aggregating semantic predications and in the same time making new inferences from the aggregated inferences.
- ▶ We also designed four measures to evaluate the performance of our technique.
- ▶ Preliminary experimental results are promising.
- ▶ We can use this technique as complementary to the semantic MEDLINE summarization.

# Acknowledgment

I would like to thank

- ▶ Dr. Olivier Bodenreider
- ▶ Dr. Paul Fontelo
- ▶ Dr. Marcelo Fizman
- ▶ Dr. Thomas Rindflesch
- ▶ Dr. McDonald
- ▶ Lister Hill Center