# Comparing the Effects of Two Semantic Terminology Models on Classification of Clinical Notes: A Study of Heart Murmur Findings

**Guoqian Jiang, Ph.D.**

**and Christopher G. Chute, M.D., Dr. P.H.**

**Division of Biomedical Informatics,**

**Mayo Clinic College of Medicine, Rochester, MN**

June 02, 2008

KR-MED 2008, Phoenix, AZ

# Introduction

- There is an emerging convergence of approaches toward the use of a concept-based clinical terminology with an underlying formal semantic terminology model (STM).

# Two types of STMs

- The existing SNOMED CT model
  - Mainly pre-coordination oriented
  - With support of post-coordination
  - Example:
    - [Hypophysectomy (52699005) ]
      - +[transfrontal approach (65519007)]
- A template model (for a specific domain)
  - Mainly post-coordination oriented
  - A list of atomic terms is organized within a semantic structure
  - Example:
    - 7-Axis model in ICNP (nursing domain)

# Problem:

- STM:
  - to support capturing structured clinical information that is crucial for computer programs
  - The structured recording has potential to improve information retrieval
- However, the functional difference in retrieval performance has not been clearly demonstrated between these two different semantic terminology models.

# Objectives of the study

- To annotate the real clinical notes using the two schemas established from two models
- To compare and evaluate the effects of two models on classification of the clinical notes
  - A study for the domain of heart murmur findings

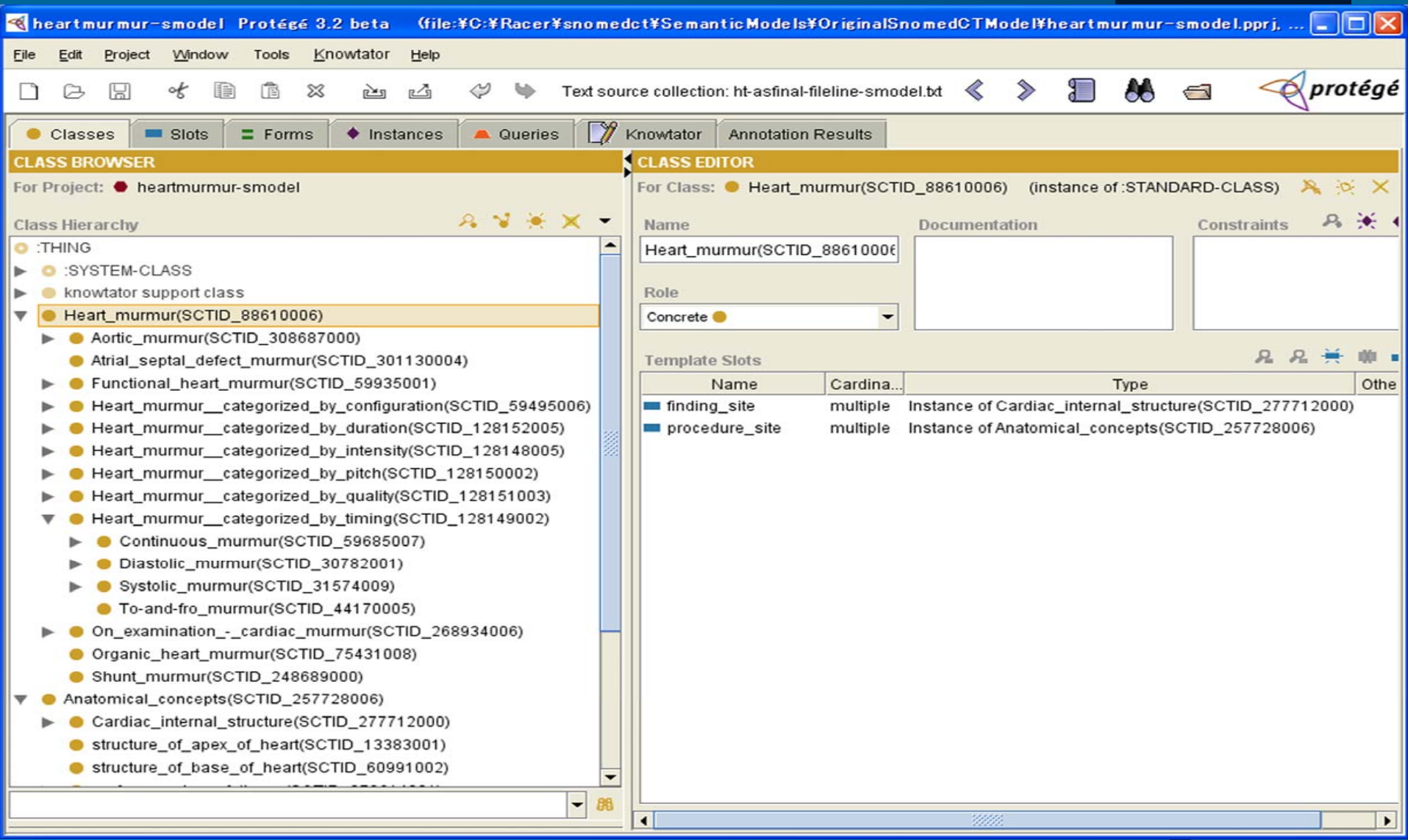
# Methods & Materials

- Defining the annotation schemas
- Preparing clinical notes corpus
- Annotation software and annotators
- Coding for machine learning classification
- Outcome measures and statistical analysis

# SNOMED CT Model Schema

- For the S-Model, we established a schema by extracting concept trees from the existing sub-hierarchy of heart murmur findings in January 2006 version of SNOMED CT.
- One root concept is “heart murmur (88610006)” which includes 86 sub-concepts of pre-coordinated terms of heart murmur findings.
- The other root concept is “anatomical concepts (257728006)” .

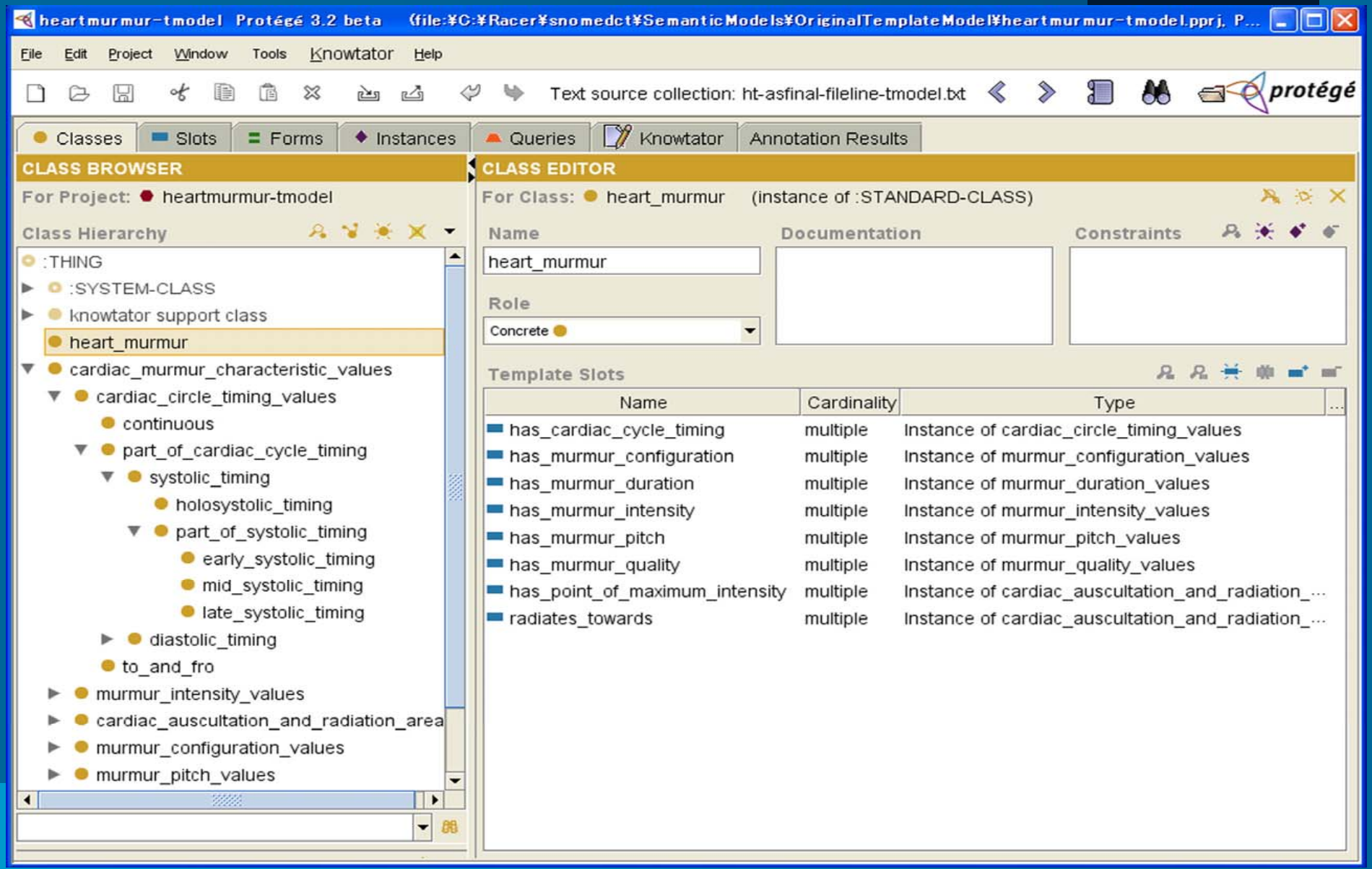
# SNOMED CT Model Schema



# Template Model Schema

- For the T-Model, a schema was established from a concept-dependent attributes model published in a recent JAMIA paper of Green, et al.
- In this schema, one root concept is “heart murmur” which had eight semantic attributes.
- We adopted the model attributes directly from Green’s model, as well as their values (kindly provided by Green, interpersonal communication).

# Template Model Schema



# Preparing clinical notes

- We extracted notes with these criteria from Mayo repository in an automatic way:
  - 1) created between January 1, 2005 to January 31, 2005;
  - 2) Having a heart murmur description in *Physical Examination* section;
  - 3) age ≥ 21;
  - 4) Having a Hospital International Classification of Disease Adaptation (HICDA) code of the Heart Valvular Disease,
  - and 5) removing patients with a code for status prosthetic valve or complication of a prosthetic valve.

# Sampling (matched by Gender & Age)

AS Non-AS AS Non-AS

| Age | Male | Control | Female | Control | Total |
|---|---|---|---|---|---|
| 21-30 | 1 | 2 | 0 | 0 | 3 |
| 31-40 | 0 | 0 | 0 | 0 | 0 |
| 41-50 | 0 | 0 | 2 | 4 | 6 |
| 51-60 | 4 | 8 | 0 | 0 | 12 |
| 61-70 | 7 | 14 | 5 | 10 | 36 |
| 71-80 | 26 | 52 | 7 | 14 | 99 |
| 81-90 | 24 | 48 | 21 | 42 | 135 |
| 91- | 2 | 4 | 4 | 8 | 18 |
| Total | 64 | 128 | 39 | 78 | 309 |

# Annotation using Knowtator

- A general purpose text annotation tool, Knowtator, was used to map text contents to our schema.
- One author (GJ) performed the annotation task and then the other author (CGC) verified the annotations for 10% of all documents. Differences were mutually adjudicated and lessons generalized to the remaining 90% of cases.

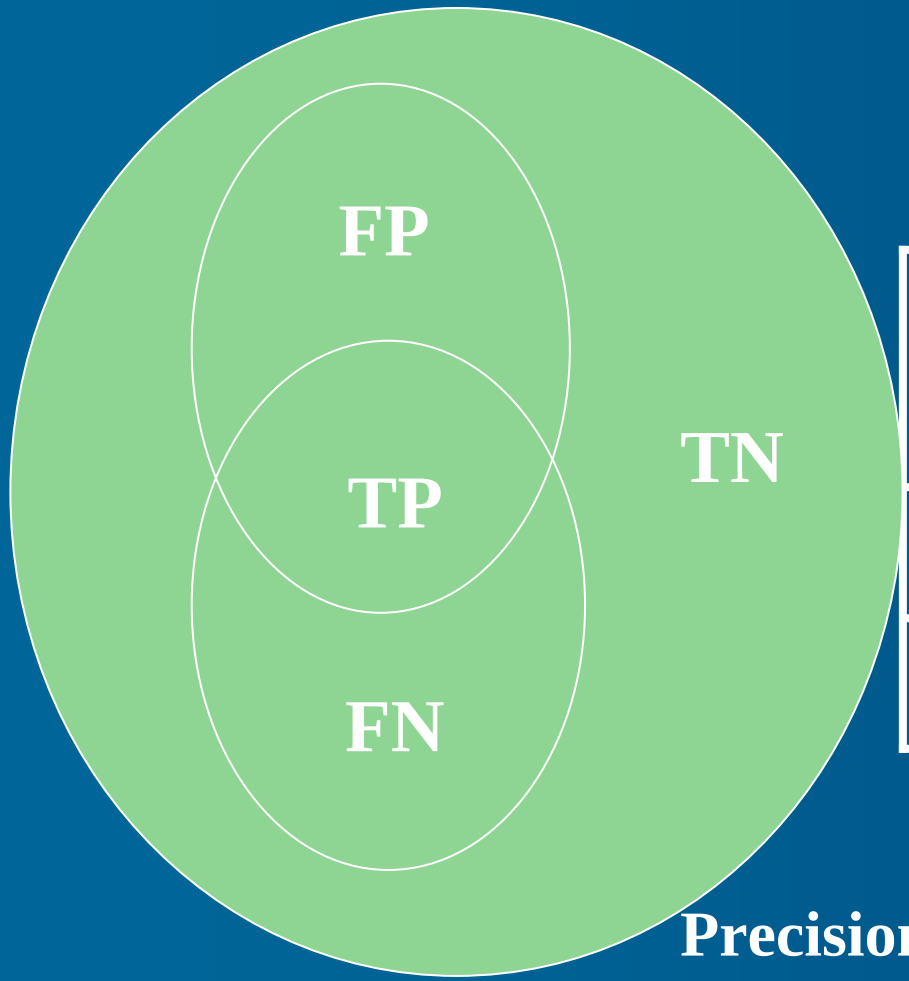
# Coding for machine learning classification

- We coded the annotated corpora for retrieval using a machine learning classification algorithm.
  - The target category of the classification is binary, i.e. aortic stenosis (AS) or non-AS.
  - In other words, the goal of the classification is to predict whether a document with a heart murmur description belongs to AS category or not.
  - The annotations of each document were used as the predictive features and coded as binary.
- We used a Weka implementation of the decision tree (J4.8).

# Outcome measures and statistical analysis

- For the annotation task
  - the description completeness
    - The annotators were asked to judge whether the heart murmur descriptions of each document could be described completely through using the schema of a model while they performed annotation task. If they judged a document as “incomplete”, they indicated a reason for the judgment.

# Standard evaluation measures




| | AS | Non-AS |
|---|---|---|
| Hit | TP | FP |
| Non-Hit | FN | TN |

**Precision = TP/(TP+FP)**
**Recall = TP/(TP+FN)**
**Accuracy = (TP+TN)/(TP+FP+FN+TN)**
**F-score = 1/(α/P+(1- α)/R)**
**α=0.5 indicating equal weight of P & R**

# Four datasets

- SM: original SNOMED CT Model
- TM1: original Green's Template Model
- TM2:completing the values for all eight semantic attributes  whenever a description appearing in the clinical notes corpus did not have a corresponding value in TM1.
- TM3: adding a new semantic attribute "has an inference to specific murmurs or etiological mentions" to the root concept "heart murmur" and also completing its corresponding values from those descriptions appearing in the corpus.

# Outcome measures and statistical analysis

- Ten-fold cross validation for retrieval was performed 10 separate times over all four datasets.
- The paired t-test was performed to test the statistical significance of performance measures between the dataset of S-Model and three datasets of T-Model.

# Results: For annotations

| | Total No. of annotations | No. of annotations per document |
|---|---|---|
| S-Model | 995 | 3.2 |
| T-Model | 1377 | 4.5 |

# Description completeness

| | No. of incomplete cases | Incompleteness | Completeness |
|---|---|---|---|
| S-Model (SM) | 88 | 28% | 82% |
| T-Model (TM1) | 201 | 65% | 35% |

# Reasons for the incompleteness of four datasets

| | SM | TM1 | TM2 | TM3 |
|---|---|---|---|---|
| Auscultation area | 1 | 78 | 0 | 0 |
| Radiation | 47 | 47 | 0 | 0 |
| Configuration | 8 | 8 | 0 | 0 |
| Quality | 7 | 5 | 0 | 0 |
| *Specific murmurs* | | | | |
| Ejection murmur | 0 | 107 | 107 | 0 |
| Regurgitant murmur | 3 | 3 | 3 | 0 |
| Flow murmur | 2 | 2 | 2 | 0 |
| *Etiological mentions* | | | | |
| Aortic valve related | 19 | 25 | 25 | 0 |
| Mitral valve related | 4 | 4 | 4 | 0 |
| Pulmonary valve related | 1 | 1 | 1 | 0 |
| Septal defect | 1 | 1 | 1 | 0 |

# Results of the evaluation metrics of the four datasets

| | Precision (mean±sd) | Recall (mean±sd) | F-score (mean±sd) | Accuracy (mean±sd) |
|---|---|---|---|---|
| SM | 74.2% ±13.7% | 59.4% ±15.6% | 64.5% ±12.7% | 79.0% ±6.1% |
| TM1 | 67.5% ±14.9% | *44.6% ±13.8% | *52.1% ±11.5% | 73.6% ±5.4% |
| TM2 | 71.0% ±14.0% | 53.2% ±18.9% | 59.0% ±15.3% | 76.9% ±6.8% |
| TM3 | 80.0% ±12.2% | *69.8% ±14.6% | 73.5% ±10.4% | *83.6% ±5.8% |

*$p < 0.05$ (paired t-test)

# Discussion

- We found that the description completeness of the S-Model was better than the original T-Model with original value set, correspondingly the performance of the S-Model on classification was also better.
- The extensions of T-Model that improved the description completeness, did improve its performance on classification.
- We clearly demonstrated that the domain coverage of a terminology model was directly correlated with its performance on classification; this is not surprising.

# Discussion

- We could see that the effect of a terminology model on its functional performance in a specific domain mainly depends on its ability to represent the contents of the domain.
- In other words, the key issue for a terminology model is how to achieve complete domain coverage. If two different terminology models could represent the contents of a domain to achieve the same coverage, their performances on classification should have no difference.

- An important conclusion of these experiments is that a template-style terminology model more readily identifies gaps in coverage, and facilitates their completion for retrieval tasks.

# Merits of Knowtator

- The first merit is that Knowtator uses the Protégé ontology editing environment to build the annotation schema. The frame-based knowledge representation system provides a flexible and expressive way to efficiently make schemas of the two model types in this study.
- The second merit is that Knowtator provides visualization of annotations, making the annotation task and confirmation process simple and efficient.
- The third merit is that the Java API of the system, which supports the annotation query that exports our coding of annotations to a classifier format automatically.

# Limitations

- We regard the evaluation in this study in its comparative context across models; absolute measures of precision and recall are subject to factors beyond the scope of this study.
- The annotations of clinical notes depends entirely on what clinicians decide to document for each patient, who they may or may not know has AS at the time. The local culture around documentation seems possible that these findings could be different on another corpus.
- We only collected a relatively small size of clinical notes corpus given that the intensive annotation tasks were required.

# Conclusion

- In conclusion, the domain coverage of the two models and their performance on classification clearly  differ when applied to real clinical notes.
- Our approach provides an effective framework to evaluate the coverage and functional performance of the semantic terminology models in a specific domain for potential improvement.

# Acknowledgements


- This study is partly supported by NIH R01 LM07319.
- The authors would like to thank Philip V. Ogren, Serguei V.S. Pakhomov, Guergana K. Savova, Pauline Funk and James D. Buntrock for their support.