

Board of Scientific Counselors  
April 8, 2010



# Quality Assurance in Biomedical Terminologies and Ontologies



*Olivier Bodenreider, MD, PhD*

Lister Hill National Center  
for Biomedical Communications  
Bethesda, Maryland - USA

# Outline

- ◆ Analytical framework for quality assurance research
- ◆ Overview of 36 auditing studies
- ◆ Detailed presentation of 4 studies



# Medical Ontology Research

- ◆ Objective: To develop methods whereby ontologies can be acquired from existing resources and validated against other knowledge sources, including the Unified Medical Language System (UMLS)
- ◆ Quality assurance
  - In and of itself
  - As part of other projects (e.g., alignment)

# Motivation

- ◆ To improve the quality of biomedical terminologies and ontologies
- ◆ Approach
  - Principled
  - Automated
  - Scalable
- ◆ Report to the developers
- ◆ Share our methods



# Ontology vs. other artifacts

- ◆ Ontology
  - Defining types of things and their relations
- ◆ Terminology
  - Naming things in a domain
- ◆ Thesaurus
  - Organizing things for a given purpose
- ◆ Classification
  - Placing things into (arbitrary) classes
- ◆ Knowledge bases
  - Assertional vs. definitional knowledge



# Ontology vs. other artifacts (revisited)

- ◆ Lexical and terminological resources
  - Mostly collections of names for biomedical entities
  - Often have some kind of hierarchical organization (e.g., relations)
- ◆ Ontological resources
  - Mostly collections of relations among biomedical entities
  - Sometimes also collect names

*In practice: “Terminologies and Ontologies”*



# Analytical framework for quality assurance research

# Analytical framework for QA research

- ◆ Special issue of JBI on “Auditing terminologies”
- ◆ Zhu et al. JBI 2009 review article
- ◆ Analytical framework
  - What is analyzed
  - Which source of knowledge
  - Which method

Zhu, X., J.W. Fan, D.M. Baorto, C. Weng, and J.J. Cimino, *A review of auditing methods applied to the content of controlled biomedical terminologies. J Biomed Inform, 2009. 42(3): p. 413-25.*



# What is analyzed

- ◆ Term/concept
  - Coverage (missing terms/concepts)
  - Wrong synonymy relation
  - Redundant concepts
- ◆ Relation
  - Missing relations
  - Inaccurate relations
- ◆ Categorization
  - Wrong categorization

# Which source of knowledge

- ◆ Intrinsic – the terminology itself
  - Terms/Concepts
  - Relations
  - Categorization
- ◆ Extrinsic – external resources
  - Corpus
    - Text corpus – identify terms in text
    - Annotation corpus – identify relations from co-occurring terms
  - Mapping

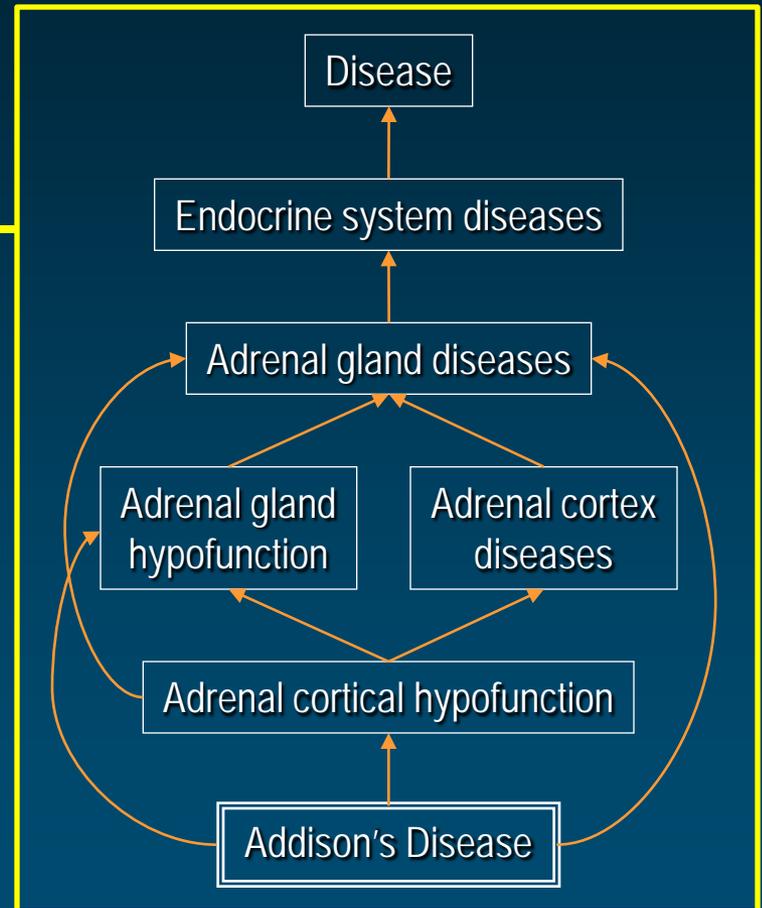
# Which method Main categories

- ◆ Lexical
  - Properties of the term
- ◆ Structural
  - Properties of the organizational structure (relations)
- ◆ Semantic
  - Semantic properties of the concept (semantic type)
- ◆ Statistical
  - Associations among entities

	Ischemic enteritis
Acute	Ischemic enteritis
Chronic	Ischemic enteritis
Modifier	Head

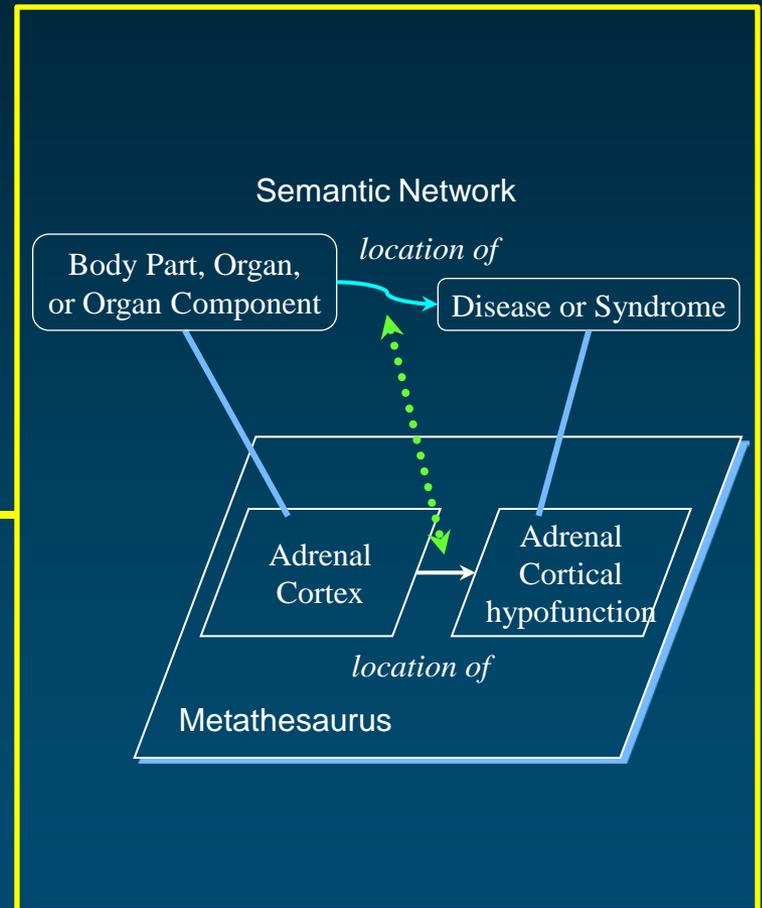
# Which method Main categories

- ◆ Lexical
  - Properties of the term
- ◆ Structural
  - Properties of the organizational structure (relations)
- ◆ Semantic
  - Semantic properties of the concept (semantic type)
- ◆ Statistical
  - Associations among entities



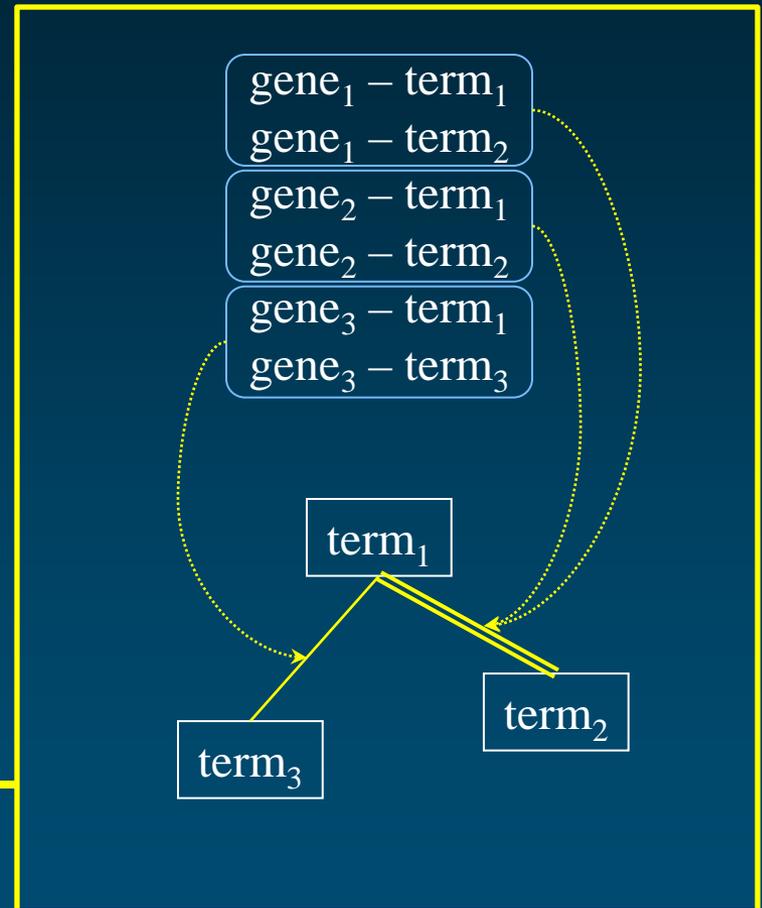
# Which method Main categories

- ◆ Lexical
  - Properties of the term
- ◆ Structural
  - Properties of the organizational structure (relations)
- ◆ Semantic
  - Semantic properties of the concept (semantic type)
- ◆ Statistical
  - Associations among entities



# Which method Main categories

- ◆ Lexical
  - Properties of the term
- ◆ Structural
  - Properties of the organizational structure (relations)
- ◆ Semantic
  - Semantic properties of the concept (semantic type)
- ◆ Statistical
  - Associations among entities



# Which method Additional methods

- ◆ Compliance with ontological principles
  - Operational definitions
- ◆ Comparative
  - Comparisons between ontologies (mapping)
- ◆ Transformative
  - Representation formalism
- ◆ Use in an application

*“Each concept, except for the root, must have (at least) one parent concept”*

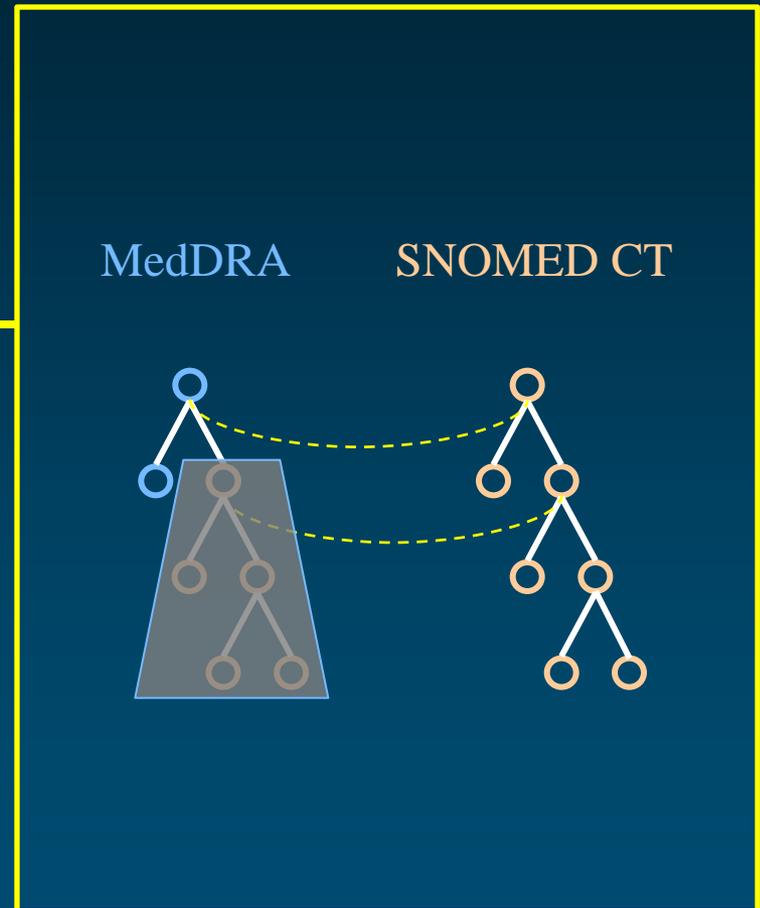


# Which method Additional methods

- ◆ Compliance with ontological principles
  - Operational definitions
- ◆ Comparative

---

  - Comparisons between ontologies (mapping)
- ◆ Transformative
  - Representation formalism
- ◆ Use in an application



# Which method Additional methods

### Heart in OWL DL

```
<owl:Class rdf:ID="Heart">
  <owl:equivalentClass>
    <owl:Class>
      <owl:intersectionOf rdf:parseType="Collection">
        <owl:Class rdf:about=
          "#Organ_with_cavitated_organ_parts"/>
        <owl:Restriction>
          <owl:onProperty
            rdf:resource="#constitutional_part" />
          <owl:someValuesFrom
            rdf:resource="#Wall_of_heart" />
        </owl:Restriction>
        ...
      </owl:intersectionOf>
    </owl:Class>
  </owl:equivalentClass>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#bounded_by"/>
      <owl:someValuesFrom
        rdf:resource="#Surface_of_heart"/>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#arterial_supply"/>
      <owl:someValuesFrom
        rdf:resource="#Right_coronary_artery" />
    </owl:Restriction>
  ...
</owl:Class>
```

logical

### Heart in Protégé frames

```
Metaclass
(defclass Heart
  (is-a Organ_with_cavitated_organ_parts)
  ...
)

Class (Instance of Metaclass)
([Heart]
 of Organ_with_cavitated_organ_parts
 (constitutional_part
  Wall_of_heart
  Cavity_of_left_atrium
  Cavity_of_right_ventricle
  Cavity_of_left_ventricle
  Right_coronary_artery
  Left_coronary_artery
  ...
 (bounded_by
  Surface_of_heart)
 (arterial_supply
  Right_coronary_artery
  Left_coronary_artery)
  ...
)
```

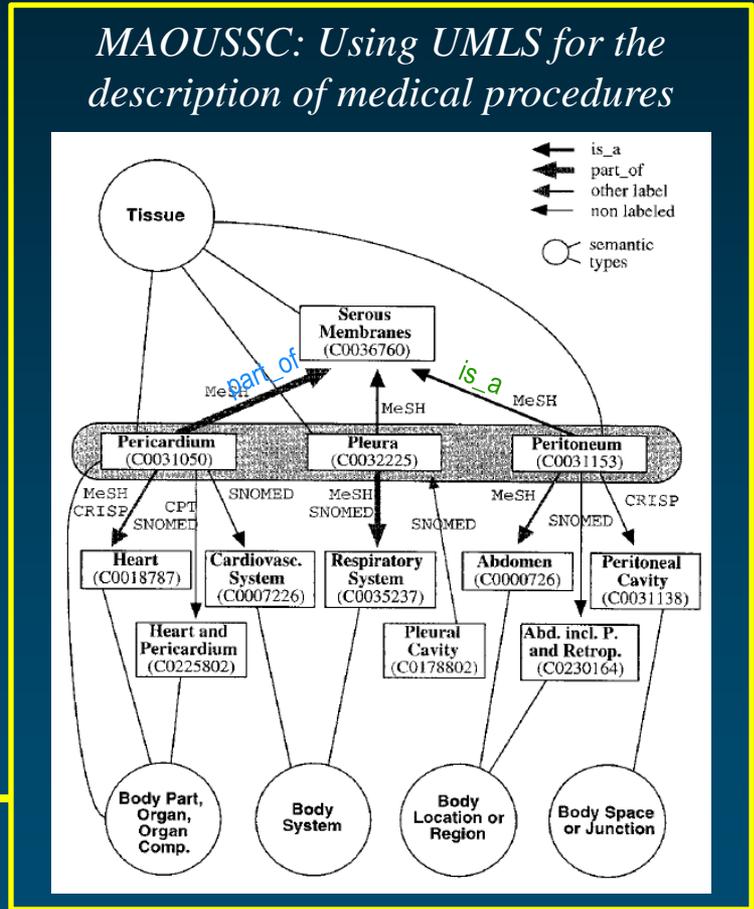


n



# Which method Additional methods

- ◆ Compliance with ontological principles
  - Operational definitions
- ◆ Comparative
  - Comparisons between ontologies (mapping)
- ◆ Transformative
  - Representation formalism
- ◆ Use in an application



# Overview of 36 research studies

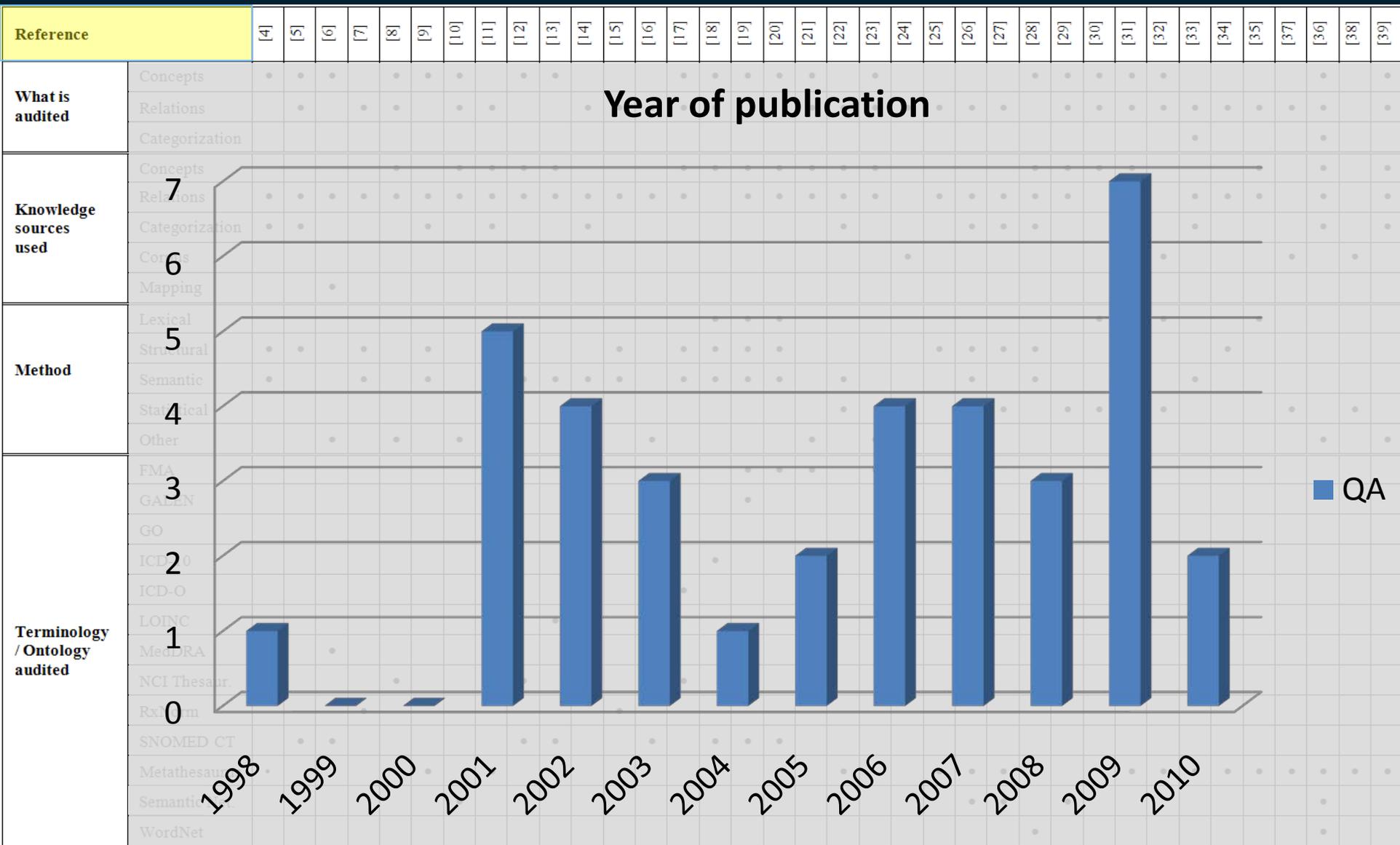
# Overview

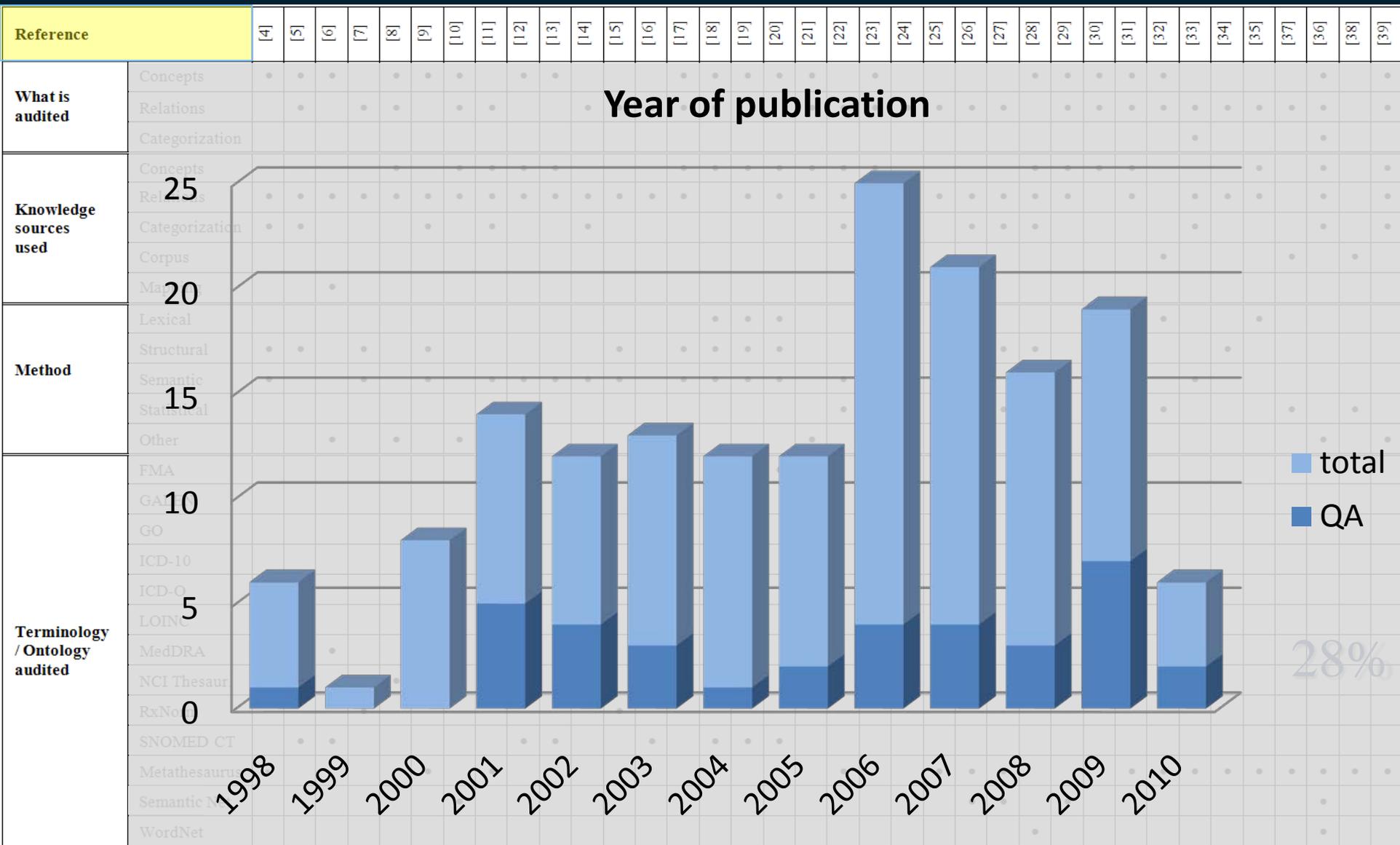
- ◆ 1998-2010
- ◆ 36 research studies on QA
  - 18 as primary focus
  - 18 as an application
- ◆ 13 different terminologies and ontologies
- ◆ Using a wide range of methods
  
- ◆ Presented through the analytic framework



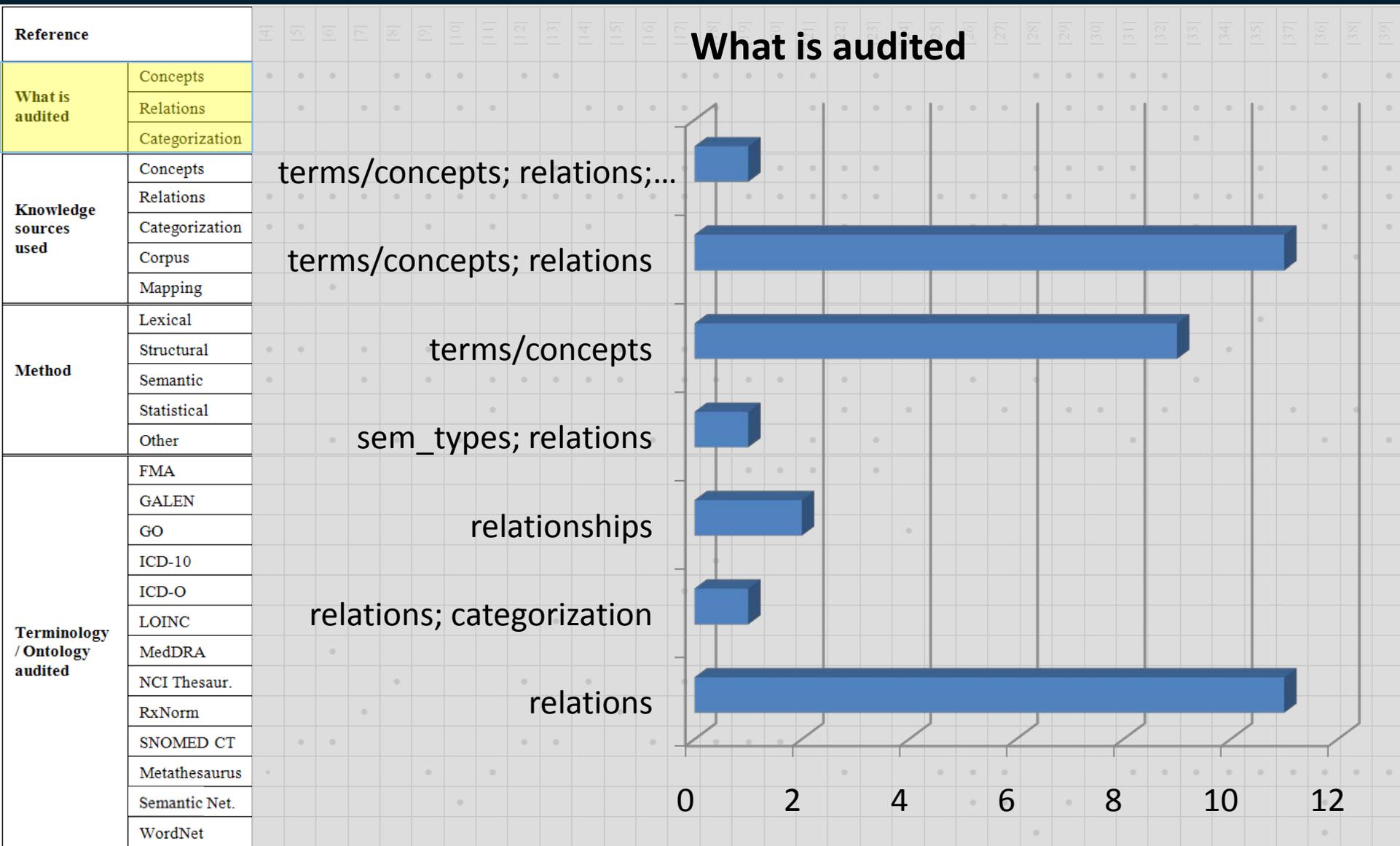
Reference		[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]	[12]	[13]	[14]	[15]	[16]	[17]	[18]	[19]	[20]	[21]	[22]	[23]	[24]	[25]	[26]	[27]	[28]	[29]	[30]	[31]	[32]	[33]	[34]	[35]	[37]	[36]	[38]	[39]		
What is audited	Concepts	•	•	•		•	•	•		•	•				•	•	•	•	•		•					•	•	•	•	•				•		•			
	Relations		•		•	•		•	•			•	•	•		•	•			•	•	•	•	•			•	•	•	•	•	•	•	•	•		•		
	Categorization																																			•		•	
Knowledge sources used	Concepts					•		•	•	•	•				•	•	•	•	•	•	•					•	•	•	•				•		•		•		
	Relations	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•		•	•	•	•	•		•	•	•	•	•		•	•	•	•		•		•		
	Categorization	•	•				•		•			•									•				•	•									•		•		
	Corpus																						•							•				•		•			
	Mapping			•																																			
Method	Lexical															•	•	•										•		•			•						
	Structural	•	•		•		•					•			•	•	•	•						•	•	•	•					•							
	Semantic	•			•		•		•	•	•	•	•		•	•	•	•			•			•		•					•								
	Statistical								•												•		•					•	•					•		•		•	
	Other			•		•		•							•					•															•		•		
Terminology / Ontology audited	FMA																•	•	•		•																		
	GALEN																•																						
	GO																					•																	
	ICD-10																																						
	ICD-O															•																							
	LOINC											•																											
	MedDRA			•																																			
	NCI Thesaur.					•					•		•		•																								
	RxNorm													•																									
	SNOMED CT		•	•							•	•			•		•	•	•																				
	Metathesaurus	•					•		•												•			•	•	•			•	•	•	•	•	•	•	•	•	•	
	Semantic Net.							•																	•	•		•								•			
	WordNet																										•									•			





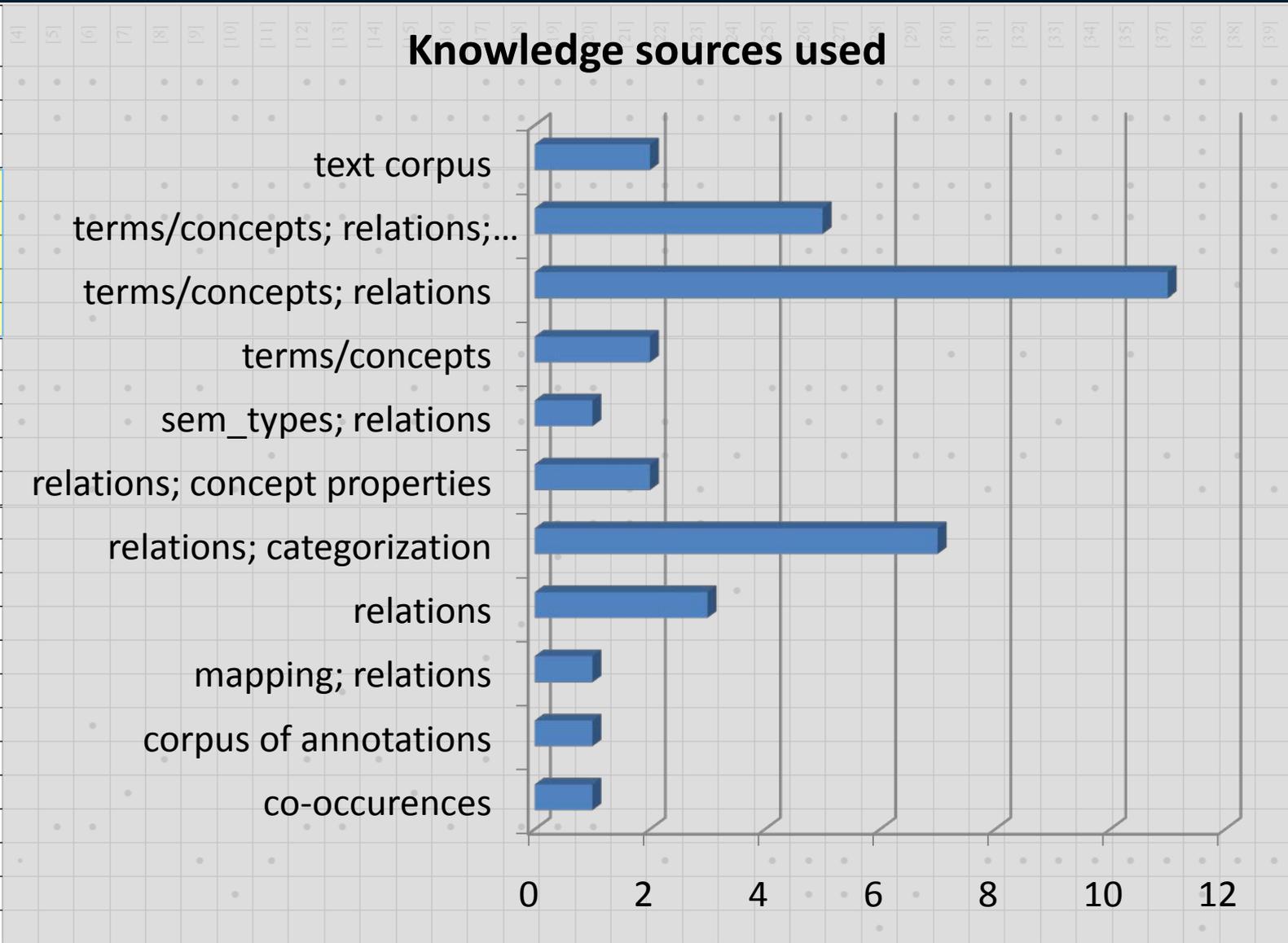




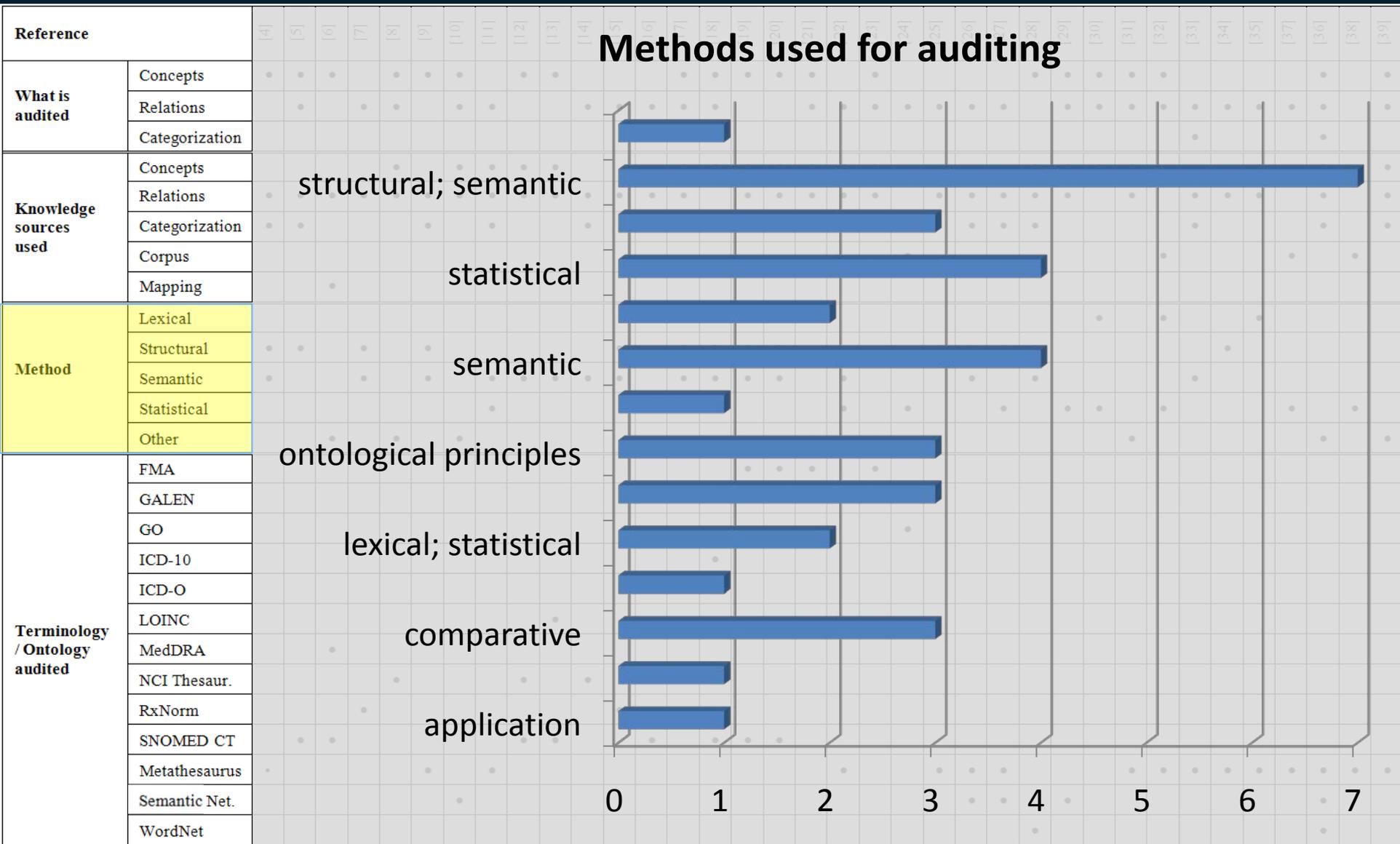


# Knowledge sources used

Reference	
What is audited	Concepts
	Relations
	Categorization
Knowledge sources used	Concepts
	Relations
	Categorization
	Corpus
	Mapping
Method	Lexical
	Structural
	Semantic
	Statistical
	Other
Terminology / Ontology audited	FMA
	GALEN
	GO
	ICD-10
	ICD-O
	LOINC
	MedDRA
	NCI Thesaur.
	RxNorm
	SNOMED CT
	Metathesaurus
	Semantic Net.
	WordNet



# Methods used for auditing



# Four examples of quality assurance research studies

# Four studies

- ◆ Identifying polysemous concepts in the UMLS
- ◆ Identifying errors in RxNorm
- ◆ Non-lexical approaches to identifying relations in the Gene Ontology
- ◆ Lexical approaches to assessing the consistency of relations in SNOMED

# *Identifying polysemous concepts in the UMLS*

Erdogan, H., E. Erdem, and O. Bodenreider, *Exploiting UMLS semantics for quality assurance purposes. Medinfo, 2010 (in press).*

# Motivation

- ◆ One error discovered in the UMLS
  - Inconsistent semantic group for the parent concept compared to the child concept
- ◆ Systematic investigation
- ◆ Semantic properties of the concepts at both ends of a hierarchical relation

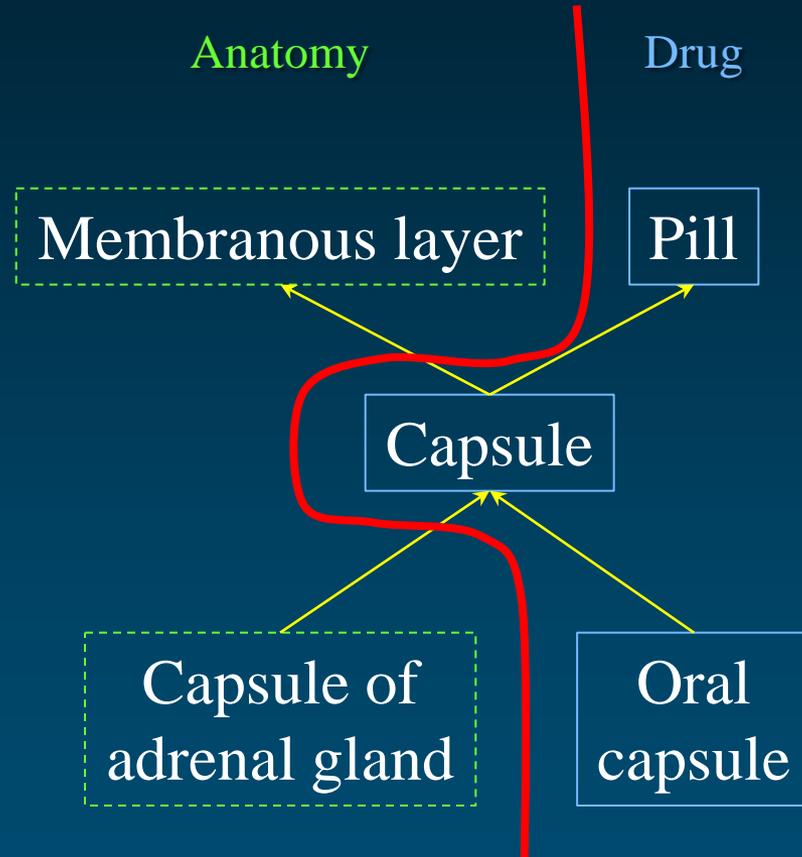
# Methods

- ◆ Check the semantic consistency
  - Semantic groups
- ◆ Pairs of hierarchically related concepts
  - 2M concepts in the UMLS
- ◆ Answer Set Programming

# Results

- ◆ 81,512 inconsistent concepts
- ◆ Most inconsistencies are not indicative of any errors
  - Navigation vs. inference
- ◆ Small number of “real” errors
  - Polysemous concepts (2 meanings in the same concept)  
= false synonymy
  - Lexically-suggested, but not caught at editing time
- ◆ Pattern for wrong synonymy

# Example

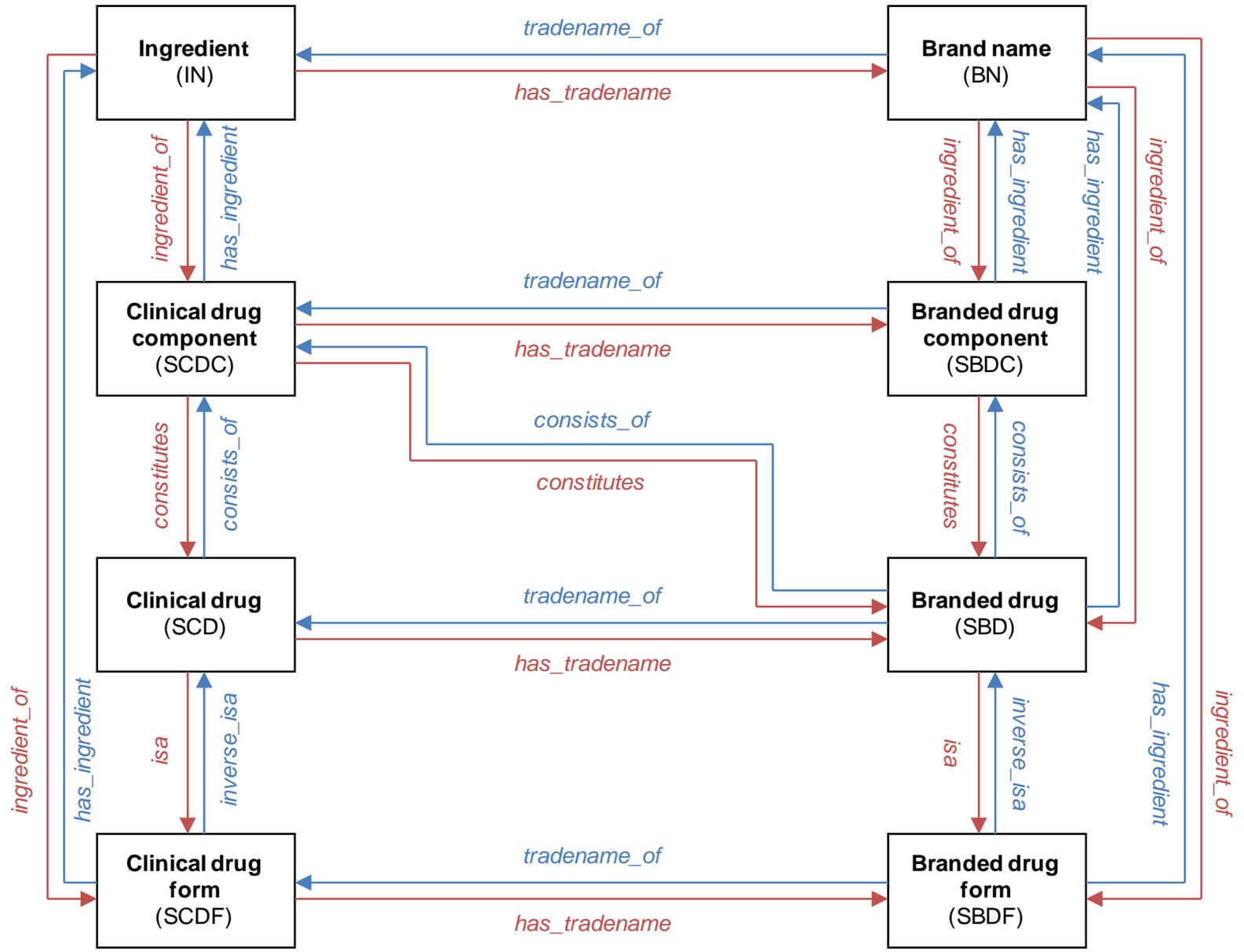


## *Identifying errors in RxNorm*

Bodenreider, O. and L.B. Peters, *A graph-based approach to auditing RxNorm. Journal of Biomedical Informatics, 2009. 42(3): p. 558-570.*

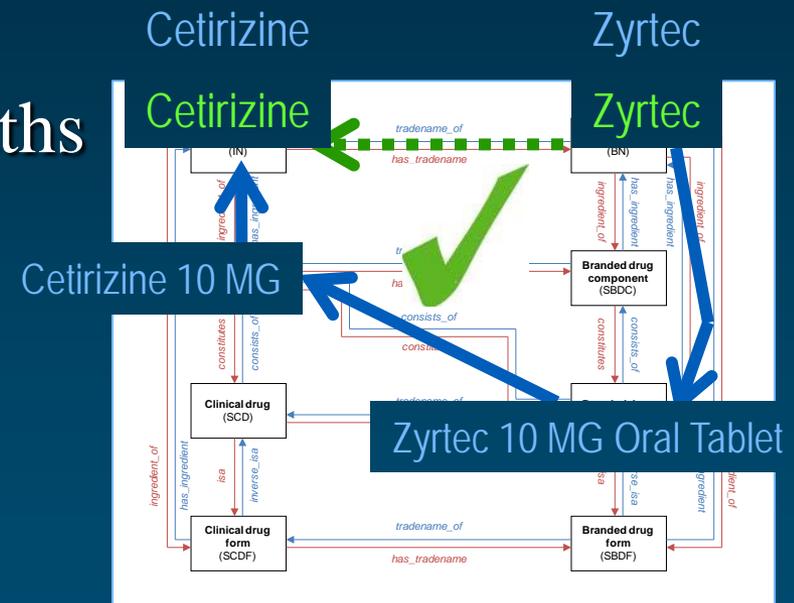
# Motivation

- ◆ Large terminology
- ◆ Relies heavily on human editors
- ◆ High quality
  
- ◆ Systematic evaluation
- ◆ Exploiting the graph structure



# Methods

- ◆ Normalize multi-ingredient drugs
- ◆ Define “meaningful” paths between 2 nodes
- ◆ Instantiate all meaningful paths
- ◆ Compare alternate paths
  - Alternate (meaningful) paths are expected to be functionally equivalent





# *Non-lexical approaches to identifying relations in the Gene Ontology*

Bodenreider, O., M. Aubry, and A. Burgun, *Non-lexical approaches to identifying associative relations in the Gene Ontology*, in *Pacific Symposium on Biocomputing 2005*, R.B. Altman, et al., Editors. 2005, World Scientific. p. 91-102.

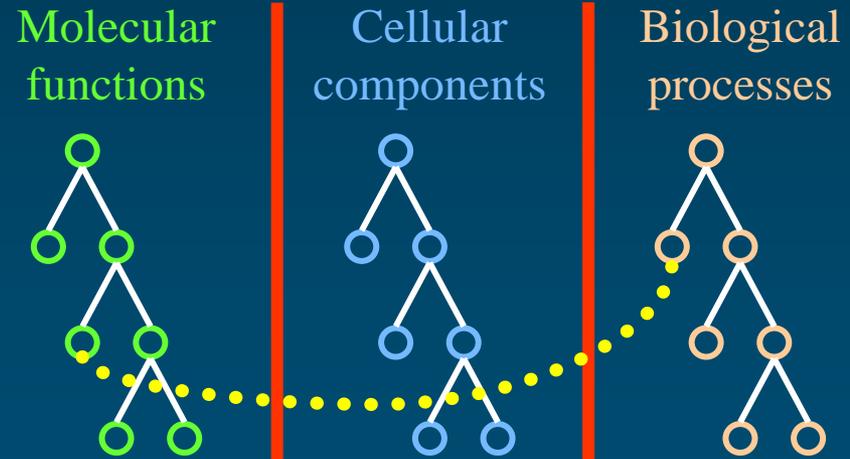
# Motivation

## ◆ Gene ontology (GO)

- Widely used for the functional annotation of gene products in many model organisms
- 3 separate hierarchies
- No relations across hierarchies

## ◆ Missing relations

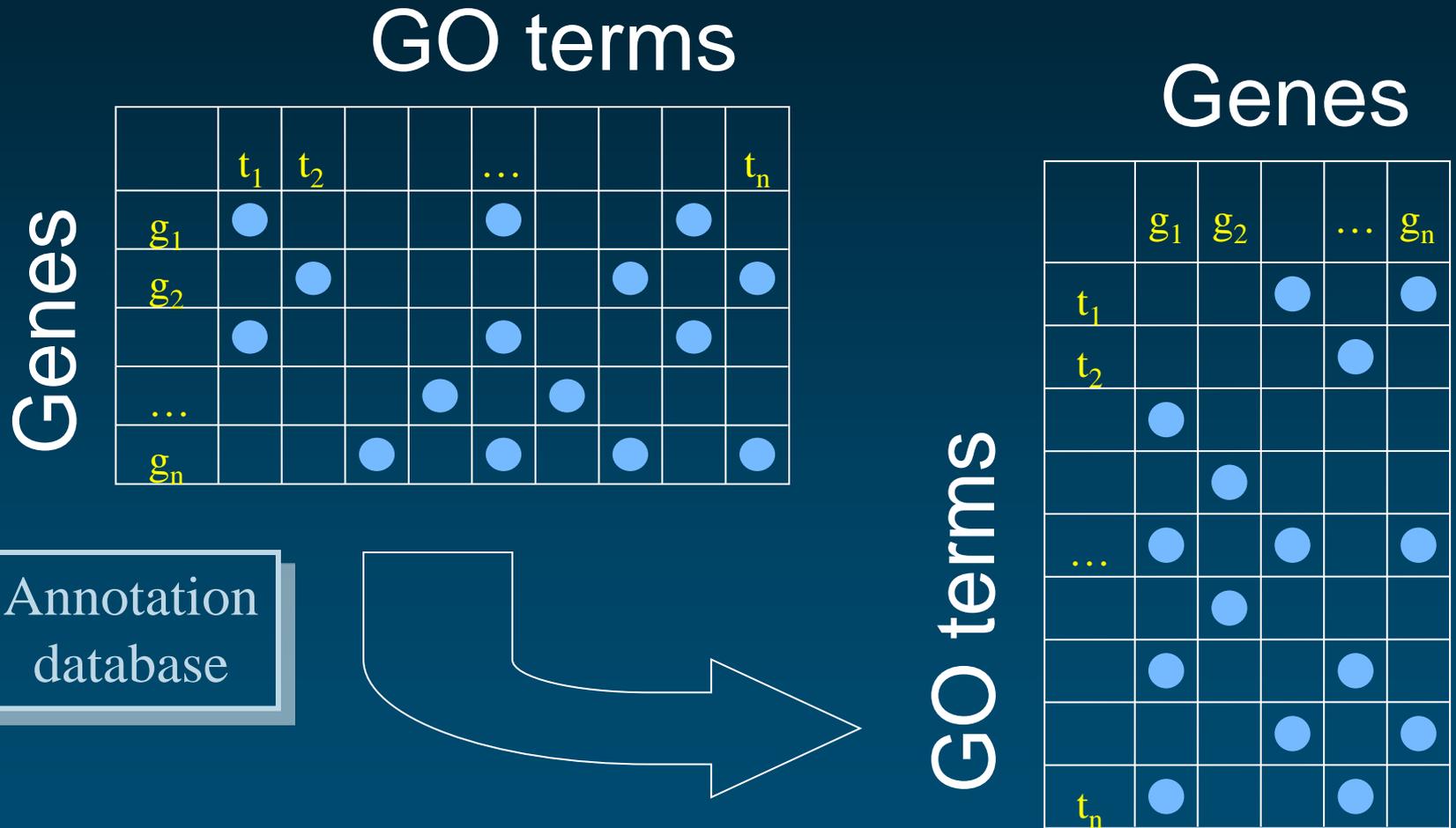
## ◆ Missing annotations



# Methods

- ◆ Lexical methods had been exploited already
- ◆ Non-lexical methods
  - Vector space model (information retrieval model)
  - Co-occurrence of annotations
  - Association rule mining

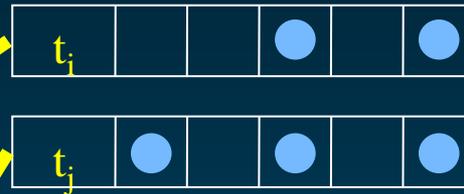
# 1 Similarity in the vector space model



# 1 Similarity in the vector space model

Genes

	$g_1$	$g_2$	...	$g_n$
$t_1$			●	●
$t_2$				●
	●			
		●		
...	●		●	●
		●		
			●	●
$t_n$	●			●



$$\text{Sim}(t_i, t_j) = \vec{t}_i \cdot \vec{t}_j$$

Similarity matrix

GO terms

	$t_1$	$t_2$	...	$t_n$
$t_1$	■	■	■	■
$t_2$	■	■	■	■
	■	■	■	■
...	■	■	■	■
	■	■	■	■
$t_n$	■	■	■	■

GO terms

GO terms



# Results Vector Space Model

	VSM	LEX
Mol Fnc-Cell Comp	499	917
Mol Fnc-Biol Proc	3057	2523
Cell Comp-Biol Proc	760	2053
<b>Total</b>	<b>4316</b>	<b>5493</b>

Mol Fnc: ice binding  
Biol Proc: response to freezing

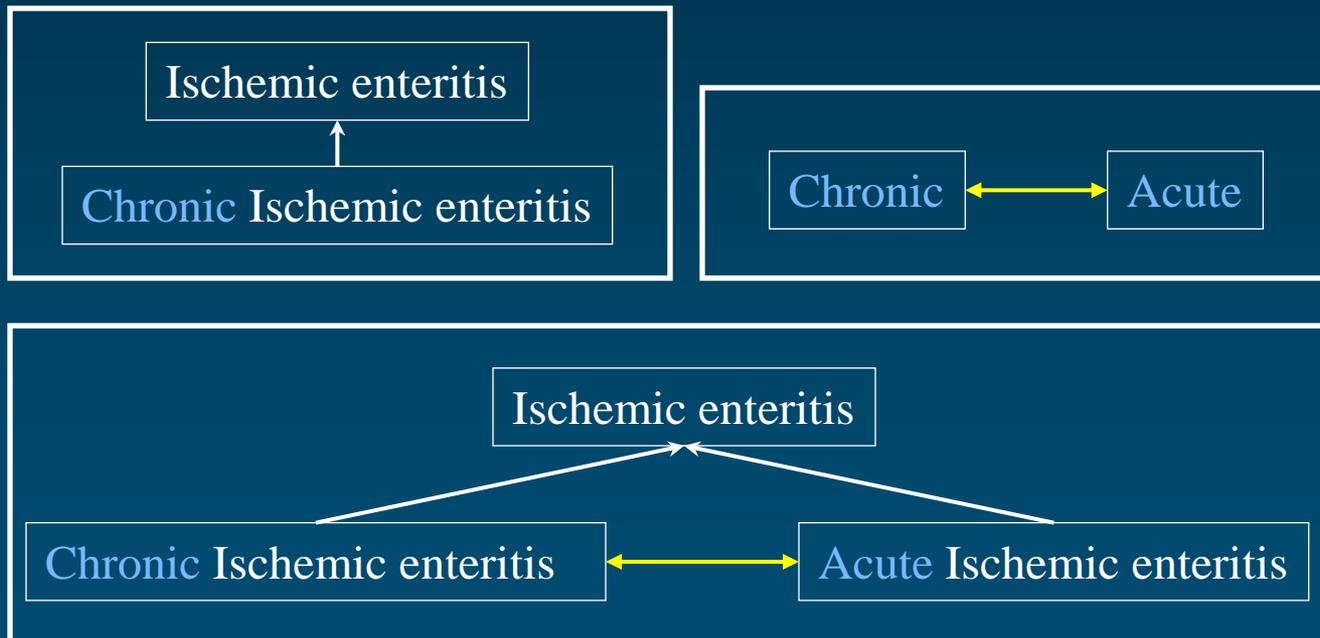


*Lexical approaches to assessing the consistency of relations in SNOMED*

Bodenreider, O., A. Burgun, and T.C. Rindflesch, *Assessing the consistency of a biomedical terminology through lexical knowledge. International Journal of Medical Informatics*, 2002. *67(1-3): p. 85-95.*

# Motivation

- ◆ Detect potential inconsistencies in SNOMED
- ◆ Based on adjectival modification
  - Frequent construct in medical terms

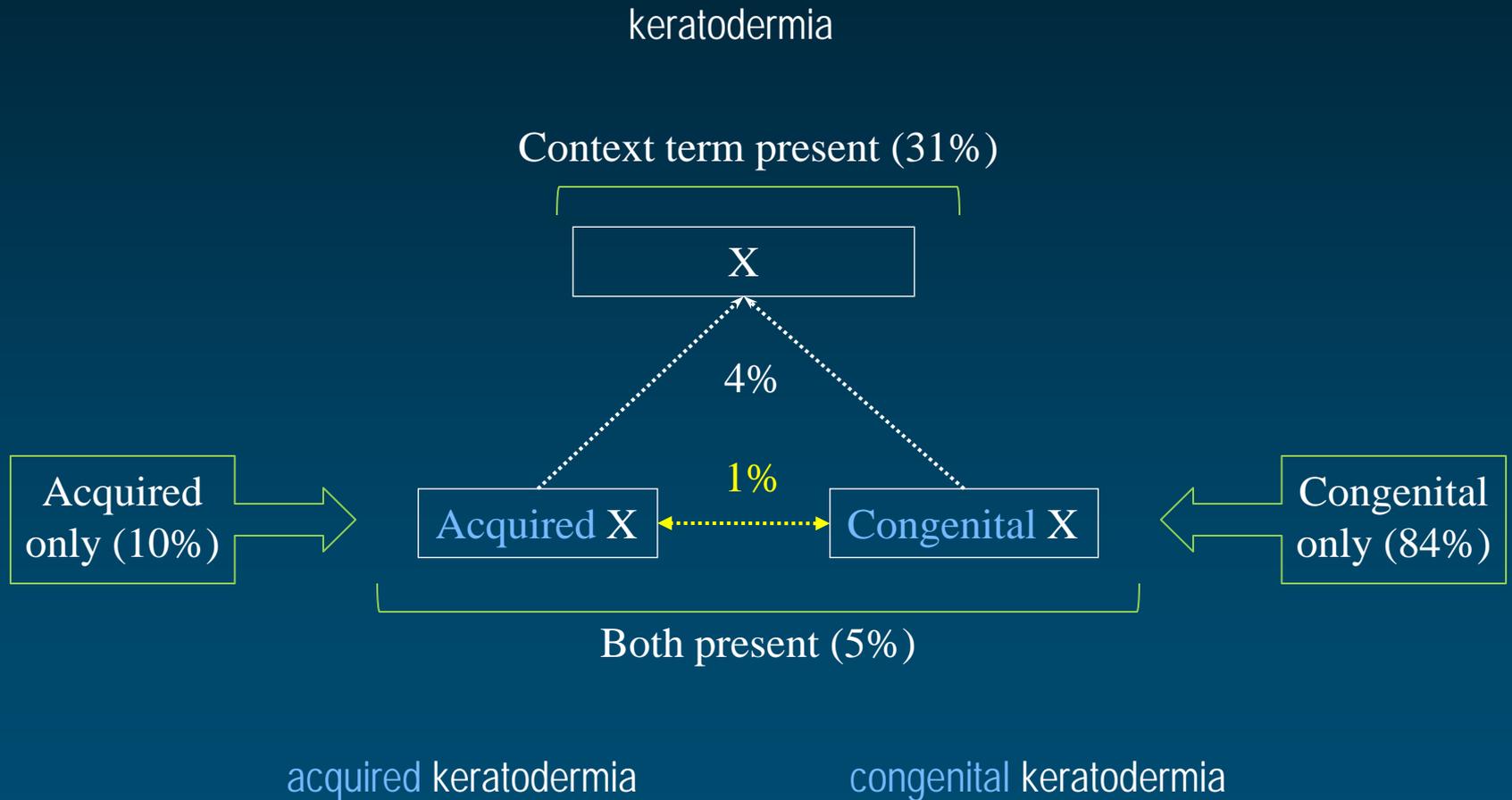


# Methods

- ◆ Identifying adjectival modifiers
- ◆ Adjectives and their contexts
- ◆ Co-occurrence of modifiers
- ◆ Generating new modified terms
- ◆ Relationship among terms associated in a pair

(acute, chronic)  
(unilateral, bilateral)  
(primary, secondary)  
(acquired, congenital)

# Results Acquired/congenital (SNOMED)



# Summary

- ◆ Investigated quality assurance in terminologies
  - Multiple approaches
  - Multiple terminologies
- ◆ Identified a limited number of errors that had defeated the quality assurance mechanisms in place in terminology development systems
- ◆ Reported to the developers
  - Fixed in some cases (FMA, RxNorm)
- ◆ Shared our methods with the scientific community

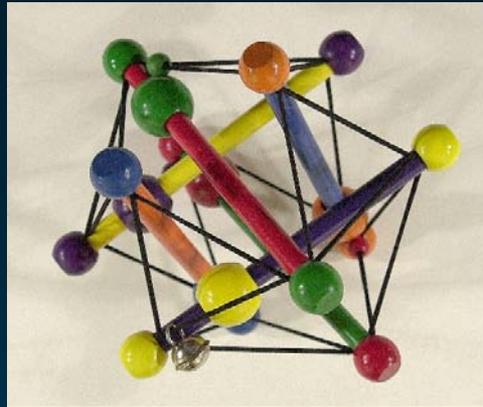
# Future work

- ◆ Develop the use of Semantic Web technologies (RDF / SPARQL and OWL) to support quality assurance
- ◆ Quality assurance “in action”
  - Investigate terminologies of clinical usefulness, (e.g., NDF-RT – clinical information about drugs)
  - Evaluate its capacity to support clinical decision
- ◆ Improve the quality of SNOMED CT through our involvement with the IHTSDO

# Acknowledgments

- ◆ NLM
  - Lee Peters
  - Kin Wah Fung
  - Lan Aronson
  - Bill Hole
  - Suresh Srinivasan
  - Tom Rindfleisch
- ◆ Harvard
  - Alexa McCray
- ◆ U. Buffalo
  - Barry Smith
  - Lowell Vizenor
- ◆ U. Utah
  - Joyce Mitchell
- ◆ NJIT
  - Duo Wei
- ◆ France
  - Fleur Mougín
  - Anita Burgun
  - Anand Kumar
  - Christine Golbreich
  - Marc Aubry
  - Genieve Botti
  - Marius Fieschi
  - Pierre Le Beux
  - François Kohler
- ◆ Germany
  - Stefan Schulz
  - Elena Beisswanger
- ◆ Netherlands
  - Erik van Mulligen
  - Lazlo van den Hoek
- ◆ PR China
  - Songmao Zhang
- ◆ Turkey
  - Halit Erdogan
  - Esra Erdem





# Medical Ontology Research

Contact: [olivier@nlm.nih.gov](mailto:olivier@nlm.nih.gov)

Web: [mor.nlm.nih.gov](http://mor.nlm.nih.gov)



*Olivier Bodenreider, MD, PhD*

Lister Hill National Center  
for Biomedical Communications  
Bethesda, Maryland - USA