



# Towards an Open Source Disease Ontology

**Prof. Warren Kibbe & Prof. Lynn Schriml**

(Northwestern University)  
wakibbe@northwestern.edu

(U. Maryland School of Medicine)  
lschriml@som.umaryland.edu

u



# Mission Statement

The Disease Ontology provides a unifying structure to map human disease knowledge between datasets such as patient records and large scale genome, sequencing and microbiome projects.



# Overview

- OBO principles
- DO Use Cases
- Medical Ontologies (just a few!)



# An Open Source Framework

## **OBO Inclusion principles include:**

- open to all for use without constraints or license
- open to modification from community input
- the ontology have a unique identifier space within OBO
- be orthogonal to other OBO ontologies
- share textual definitions in their terms
- implemented as a DAG graph (nodes with multiple parents, children relationships)
- be machine computable
- the path to the top is always true
- be based on a hierarchical structure with a single disease root
- be compatible with DAG tools such as Protégé and OBO-edit)

## **OBO Foundry Inclusion Principals include:**

- development of ontologies through collaborative efforts;
- use of unambiguously defined common relations (e.g. part\_of, is\_a)
- implemented methods to provide community input
- provide access to current and previous versions
- have a clearly bound subject-matter.

# DO Development: Use Case Driven

## Historical

- Mapping disease to billing and EMR data in clinical records

## Revision History

- DO\_V1: Designed around ICD Codes
- DO\_V2: Based on SNOMED , MeSH and UMLS
- DO\_V3: Reorganized (V2) on along more clinically relevant axes:
  - anatomical location, environment, infectious agent and aberrant process.

## Strength of the earlier versions of DO

- DO V1 was great at mapping ICD-9 data to disease
- DO V2 provided some ability to map other areas of the medical record



# DO Development: Use Case Driven

## **Future: Use Cases:**

### Future Driving Forces for DO Development

- Integrate disparate datasets, correlate disease with symptoms
- Enable the integration of disease concepts from multiple sources
- Provide 'cross-walks' between disease concepts in existing vocabularies
- Map and validate data from Electronic Medical Records to identify patients with specific disease

# Original driving Use Case



NUgene collects and stores genetic (DNA) samples along with associated healthcare information from patients of Northwestern-affiliated hospitals and clinics. It is currently the only study of its kind in Chicago and one of a few in the nation. This resource is available to scientists to conduct groundbreaking genetic research.



- The NUgene Project is a genetic banking study which collects and stores DNA samples and associated medical information from its participants
  - Medical information consists of Epic data (mostly free text) and billing/procedure data
- Problem:
  - What types of diseases do our participants have?






# Medical Ontologies

- MeSH (Medical Subject Headings)
  - Shallow graph with no direct ICD9 mapping
- NCI Thesaurus (National Cancer Institute)
  - Broad coverage, not deep outside cancer domain
  - No direct mapping to ICD9
- SNOMED (Systemized Nomenclature of Medicine)
  - Large, broad but duplicate concepts in different contexts
  - Restrictive, only free for research in the US
- ICD9/ICD10 (International Classification of Disease)
  - Poor coverage, few high level terms, confused terms



# Medical Subject Headings (MeSH)

- MeSH is widely applied to literature
- MeSH is rich from a content standpoint
- MeSH has 24,787 descriptors as of 2008 and 172,000 supplemental concepts and 97,000 entry terms (“synonyms”)
- MeSH is very flat (not deep)
- MeSH is not semantically consistent
- MeSH is not computable



# Universal Medical Language System (UMLS)

- UMLS maps to concepts and terms in LOINC, SNOMED CT, RxNorm, MeSH (more than 140 source vocab.)
- UMLS contains > 900K unique concepts
- UMLS is semantically organized
- UMLS contains cycles rather than DAGs
- <http://www.nlm.nih.gov/research/umls/>



# International Classification of Diseases (ICD)

- ICD is commonly used to report mortality and morbidity statistics
- ICD is nearly universally used to code patient encounters for billing
- ICD is not semantically organized, or rather the organization is around an encounter, not disease concepts



## Systematized Nomenclature of Human and Veterinary Medicine—Clinical Term® (SNOMED CT)

- SNOMED CT contains >350K terms and 900K descriptors
- SNOMED is content rich
- SNOMED is semantically consistent
- SNOMED is computable
- SNOMED is NOT open source
- SNOMED does not contain a disease-focused branch or view



# Disease Ontology



[Overview](#)

[Communication](#)

[Development](#)

[Downloads](#)

[Projects](#)

[Contacts](#)

## Overview

Disease Ontology is a controlled medical vocabulary developed at the Bioinformatics Core Facility in collaboration with the [NuGene Project](#) at the [Center for Genetic Medicine](#). It was designed to facilitate the mapping of diseases and associated conditions to particular medical codes such as ICD9CM, SNOMED and others. This mapping is useful in efforts like the NuGene Project because it allows requests for particular tissue type requests to be mapped quickly and with high fidelity to a set of ICD9 codes that can then be used to retrieve appropriate samples from the tissue bank. Without such a mapping, clinicians are forced to manually search through ICD9CM coding booklets to find all possible applicable codes matching their request. Given the complex organization of ICD9CM and difficulty of the manual process, codes and therefore tissue samples are often overlooked. In one sample case, an early version of the Disease Ontology doubled concept coverage while reducing the overall misclassification error percentage. Eventually we envision that the Disease Ontology can also be used to associate model organism phenotypes to human disease as well as medical record mining.

Disease Ontology is implemented as a directed acyclic graph (DAG) and utilizes the [Unified Medical Language System \(UMLS\)](#) as its immediate source vocabulary to access medical Ontologies such as ICD9CM. Using this standard, much of the process of updating the ontology can be handled by UMLS, freeing resources for clinicians to pursue more urgent tasks. For situations where the graph needs to be directly edited, the open source graph editor [DAGEdit](#) can be used. DAGEdit can readily manipulate and view the Disease Ontology because it is stored in [Open Biomedical Ontologies \(OBO\)](#) format in order to take advantage of DAGEdit and any other OBO standards compliant tools. The [screenshot](#) of the Disease Ontology was taken using DAGEdit and shows version 3 of Disease Ontology.

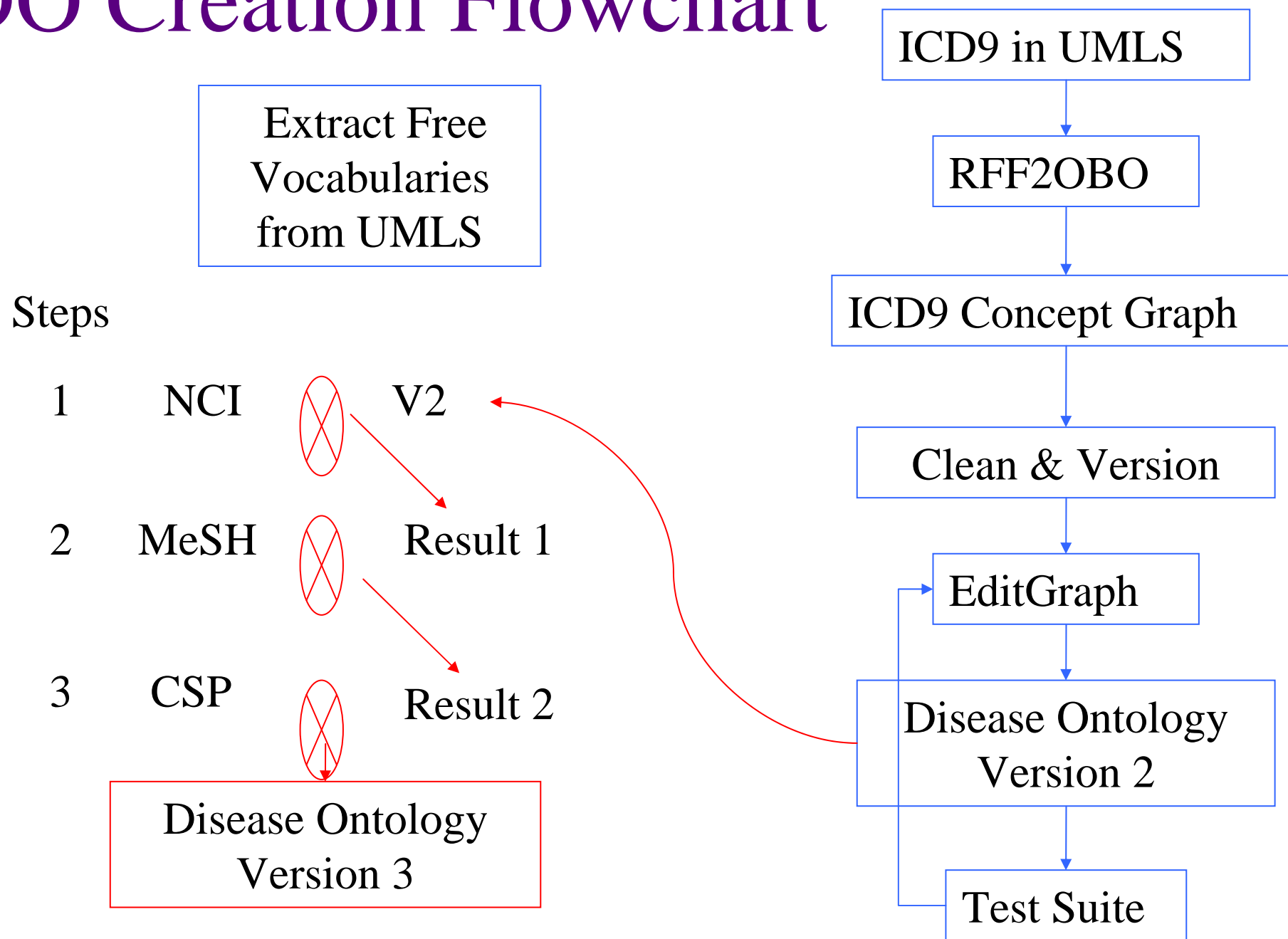
<http://diseaseontology.sourceforge.net/>



# DO Version 3

- Is an OBO Foundry ontology for the Integration of Biomedical Data
- Is inclusive of genetic, environmental and infectious diseases
- Is semantically organized and computable
- Path to the top is mostly true
- Has 12,564 terms and 21,024 branches
- Has a maximum depth of 13 and is 'node heavy in the middle', meaning many nodes are in the 6-10 node deep range

# DO Creation Flowchart





# Version 3 Graph and Statistics

- Built with 6 different vocabularies licensed as “Level 0” vocabularies under UMLS
  - NCI
  - MeSH
  - ICD9 & ICD10 (not all of ICD9 mapped, for example BP reading)
  - CSP
  - MTHICD9
- 12,564 Nodes
- 21,024 Edges
- Depth 13, node heavy in the middle ranks

# DO V3: Characteristics

- DO terms are linked to well-established, well-adopted terminologies that contain disease and disease-related concepts such as SNOMED, ICD-9 and ICD-10, MeSH, and UMLS. [in place]
- Terms in DO will be well defined, using standard references (Stedman 2005, Harrison 2003) [policy]
- The combination of a semantically computable structure and the external references to these terminologies will enable useful inference between disparate datasets using one or more of these standard terminologies to code disease [assumption]
- Open access: available on Sourceforge: <http://diseaseontology.sourceforge.net/> [in place]
- OBO Foundry commitment [policy]

# DO Mappings

<b>External Reference</b>	<b>Unique xref:DOID Mappings</b>	<b>Unique xrefs</b>
ICD-9	186278	10109
UMLS_SNOM DCT_2005_01_31_AUI	38912	38912
UMLS_NCI2004_11_17_AUI	24049	24049
UMLS_MSH2005_2005_01_17_AUI	21377	21377
UMLS_CUI	17023	17023
UMLS_ST	14674	14674
SNOMEDCT_2005_01_31	13116	13116
UMLS_ICD-9	10048	10048
NCI2004_11_17	6991	6991
UMLS_MTHICD-9_2005_AUI	3611	3611
MSH2005_2005_01_17	3502	3502
UMLS_CSP2004_AUI	2269	2269

Unique mappings in DO to ICD-9, SNOMED CT, NCI metathesaurus (EVS), MESH, UMLS and MESH terms. The compound reference names, such as UMLS\_SNOMEDCT\_2005\_01\_31\_AUI show the release of UMLS used to perform the SNOMED CT mappings. All existing mappings are to exact or closest concept match between a DO term and the external reference source.

# DO V3 To Dos

## DO Improvements:

- Ontological structure:
  - Still needs a lot of help
  - Views vs internal structure needs to be better defined  
*different communities need to view DO from very different perspectives and DO needs to easily support each viewpoint*

## Three main hurdles which face DO development:

- 1) Proper ontological structuring of DO
- 2) Mapping between source vocabulary terms for all DO terms
- 3) Providing tools and resources for community input of DO and timely updates of DO

# DO V3 Graph Issues

- Numerous composite nodes
- In appropriate inheritance created through the UMLS cross-vocabulary mapping
  - Context sensitive concepts need review
- ICD9 misery
  - Bizarre concepts are included that are better mapped to than included
    - Can not use SNOMED mappings due to licensing restrictions
    - OBO format does not represent external relationships (is\_a, part\_of, etc) for external references
- Phenotypes?



# Future Technology Development

## Ontology Tools and Technology:

Wiki Development

DO Listserv

DO Sourceforge Tracker for new term submission



# Past and Current Applications

- The primary driver for the creation of the Disease Ontology is the ability to integrate disparate datasets that contain disease concepts or concepts that can be mapped to disease.
- The Disease Ontology provides a unifying structure to map disease knowledge between datasets such as patient records and large scale genome, sequencing and microbiome projects.

**GeneRIF and NUGene studies:** DO fulfills that role in an unbiased and granular fashion by providing the key component in the arsenal of tools by providing computable relationships between disease and concepts that can be mapped to disease such as genetic associations to disease, symptoms and biological process.

**Gemina project:** DO has been utilized to annotate incidents of infectious pathogens and to provide a query and retrieval vocabulary linking disease to hosts and transmissions and outbreaks of disease. (<http://gemina.igs.umaryland.edu>)

**eMERGE Consortium's electronic medical records (EMR)** will be validated by the Disease Ontology so that we can more easily map participants to specific disease cohorts and map data coming from each EMR system to common standards.



# GeneRIF mining using DO

Using MMTx to mine GeneRIFs with DO or MeSH

DO and MeSH result in low false positive, DO has a lower false negative rate

Using MMTx to mine OMIM with DO or MeSH

Complex full sentences with compound ideas are hard to parse with standard text mining techniques – Mining OMIM did not perform as well as mining GeneRIFs

# DO and PATO

- A critical aspect in applying DO to medical literature and medical records is that ability to walk between "Signs and Symptoms" attached to a patient record or the medical literature and the disease
- PATO will describe the phenotype, possibly with a 'Signs and Symptoms' view, and DO will describe the disease concept(s) linked with those signs and symptoms
- When we map from ICD-9 to DO, for instance, we find that collections of signs and diagnoses are associated with disease, and that it is the collection, rather than a single association, that enables the inference of disease from a set of observables



# DO and model systems for disease

- DO is human-centric
- Is it enough to link disease concepts between organisms?
- What happens when the underlying mechanisms appear to have changed between the organisms?
  - Epilepsy in Dogs and Human
  - Viral Diseases
  - Rat and Mouse models and Cancer in Humans



# DO Community Engagement

## Models of Community Ontology Development

- GO
- EnvO/GAZ

DO needs to and is actively moving towards building a community of mutual invested members, most likely with multiple communities of practice growing and evolving around specific areas or viewpoints of disease.



# DO Community Engagement

## **How DO is Working Toward Engaging the Community:**

- This meeting
- More active role in OBO Foundry: attending OBO Foundry meeting in July
- Participating in other related Ontology efforts: IDO
- Medical Ontology Efforts (Medical Knowledge Research Group) Steve Roessingh, Leo Cousineau, Robert Baud
- Engaging the Neurogenomics community through Maryann Martone in BIRN

# Example Application - DO Browser

**Disease Ontology**

- Communicable Diseases [1728 patients, 2619 terms]
- Disorders of Environmental Origin [1398 patients, 1419 terms]
- Stomatognathic Diseases [339 patients, 283 terms]
- Syndrome [508 patients, 156 terms]
- Mental and behavioral problems [989 patients, 871 terms]
- Neoplasms [1468 patients, 681 terms]
- Hyperplasia [130 patients, 27 terms]
- Hemic and Lymphatic Diseases [1388 patients, 437 terms]
- Otorhinolaryngologic Diseases [2039 patients, 2043 terms]
- Skin and Connective Tissue Diseases [2213 patients, 3069 terms]
- Degenerative Disease [949 patients, 1192 terms]
- Disorder by Site [2480 patients, 8716 terms]
- Hereditary Diseases [458 patients, 113 terms]
- Digestive System Disorders [1632 patients, 1159 terms]
- Immunodeficiency and Immunosuppression Disorders [1284 patients, 422 terms]
- Deformity [758 patients, 629 terms]
- Lifestyle-related condition [627 patients, 375 terms]
- Organic brain syndrome [31 patients, 41 terms]
- Socialized Conduct Disorder [519 patients, 259 terms]
  - Socialized conduct disorder, mild degree [0 patients, 1 terms]
  - Socialized conduct disorder, severe degree [0 patients, 1 terms]
  - Undersocialized Conduct Disorder, Aggressive Type [0 patients, 5 terms]
  - Phobic anxiety disorder [5 patients, 6 terms]
  - Socialized conduct disorder, moderate degree [0 patients, 1 terms]
  - Impulse Control Disorders [0 patients, 5 terms]
  - Panic Disorder [17 patients, 2 terms]
  - Communication impairment [506 patients, 237 terms]
    - Hearing problem [158 patients, 38 terms]
    - Vision Disorders [414 patients, 196 terms]
    - Language Disorders [1 patients, 2 terms]
    - Learning Disorders [2 patients, 2 terms]
- Dependence [24 patients, 67 terms]
- Substance Withdrawal Syndrome [59 patients, 4 terms]
- Tobacco Use Disorder [90 patients, 5 terms]

ICD-9 Term(s) to Find:

Terms ANDED

34882: Hearing problem

27634: Vision Disorders

OR

Terms ANDED

BUT NOT

Terms Excluded

ICD-9 Codes (233)	ICD-9 Codes (0)	ICD-9 Codes (0)	Unique Patients 67
036.81 094.84 250.80 253.5 264.5 360.21 362.85 363.05 367			Unique Samples

Save Query

Name for Query:

Project Name:

Category:

Comments:



# Intelligent Medical Knowledge Base

Medical Knowledge  
Discovery

*The Disease Ontology is the core of the Medical Knowledge Base*

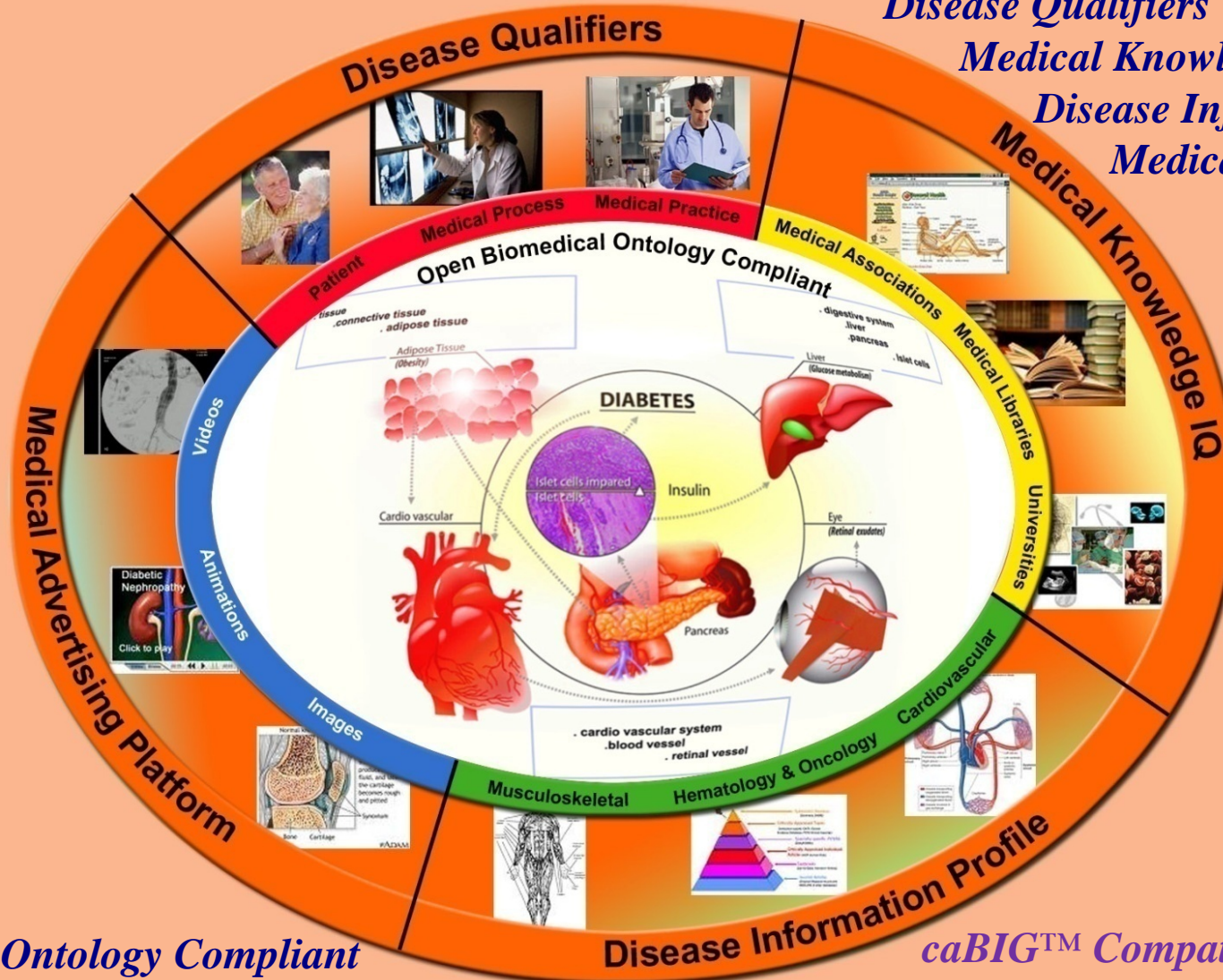
*Per disease the following are defined:*

*Disease Qualifiers*

*Medical Knowledge IQ*

*Disease Information Profile*

*Medical Advertising*



*Open Biomedical Ontology Compliant*

*Disease Information Profile*

*caBIG™ Compatible Design*



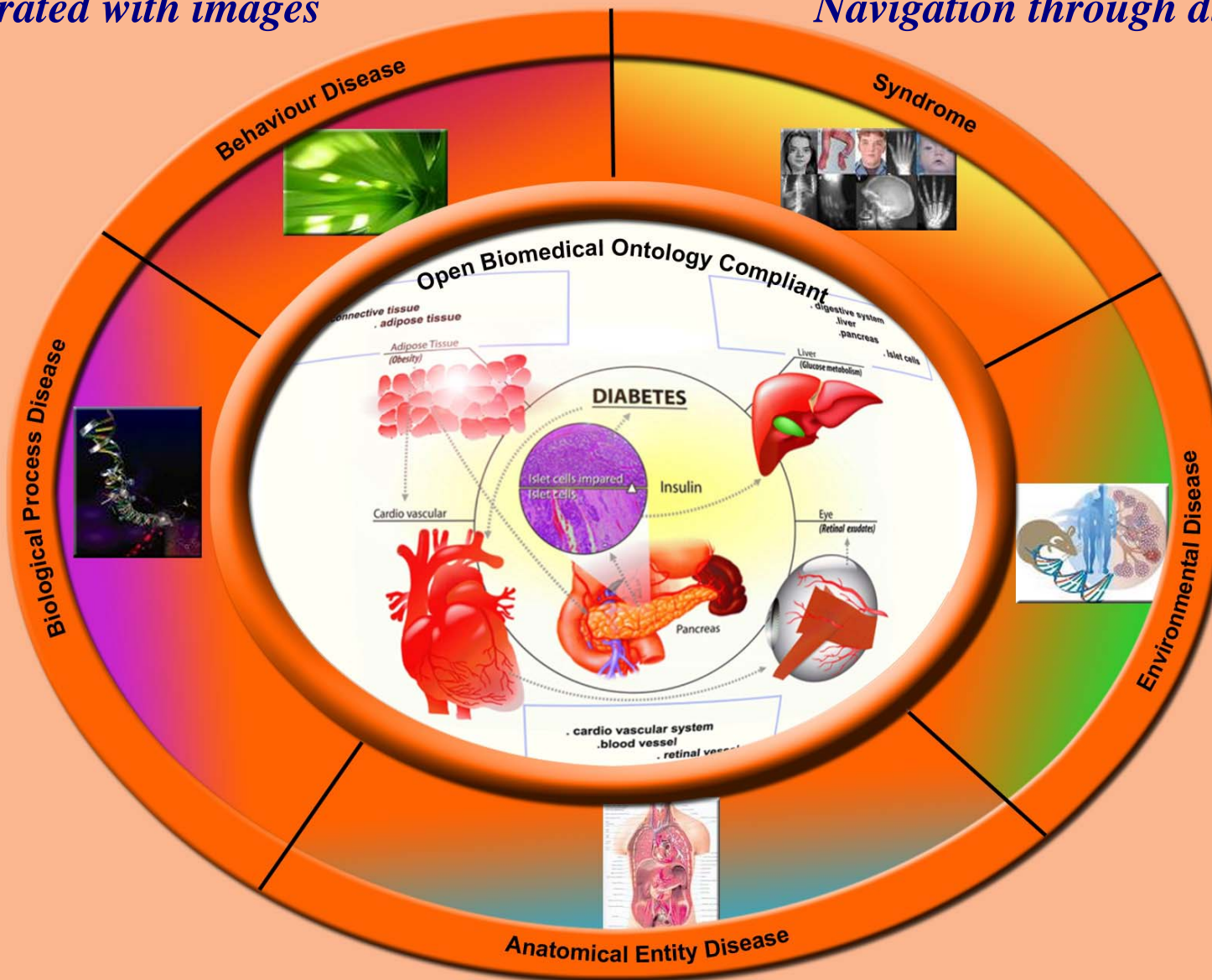
Medical Knowledge  
Discovery

# Disease Ontology

*All Medical Searches are related to the Disease Ontology*

*Diseases illustrated with images*

*Navigation through disease patterns*





Medical Knowledge  
Discovery

# Intelligent Medical Search

*Over 350 combinations of qualifiers are defined per disease*

*In 3 clicks the following are added:*

## Patient



*Child*



*Adolescent*



*Female*



*Male*



*Elderly*

## Medical Process



*Diagnostic  
Procedures*



*Diagnosis*



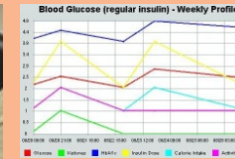
*Treatment*



*Medication*



*Follow-up*



*Tracking*



*Outcome  
Assessment*

## Medical Background



*Symptoms*



*Causes*



*Risk Factors*



*History*

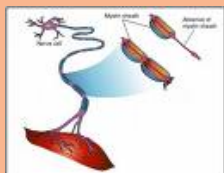


*Family History*



*Pregnancy*

## Medical Practice



*Definition*



*Alternative  
Medicine*



*Guidelines*



*Best  
Practices*



*Compliance*

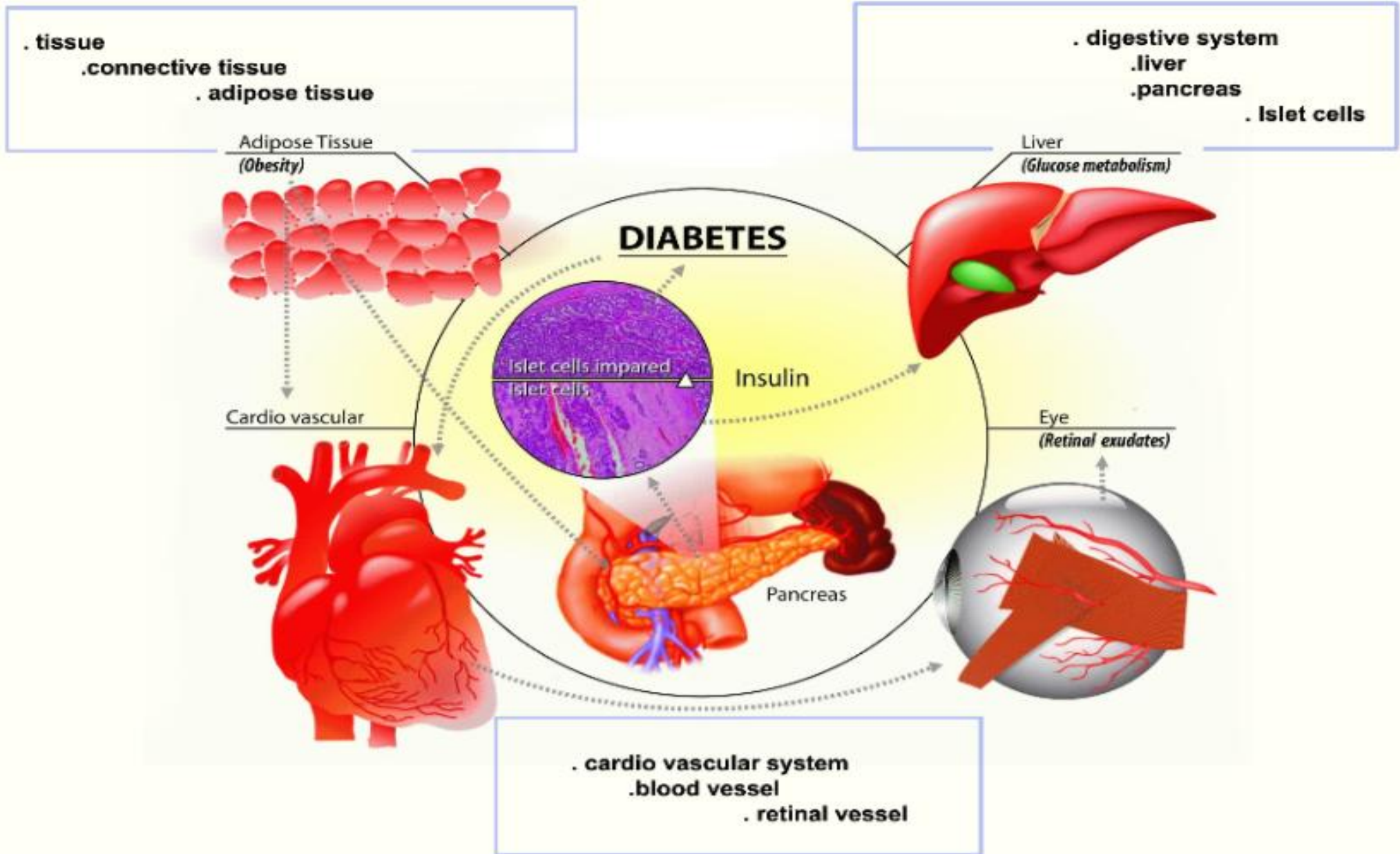


*Clinical Studies*

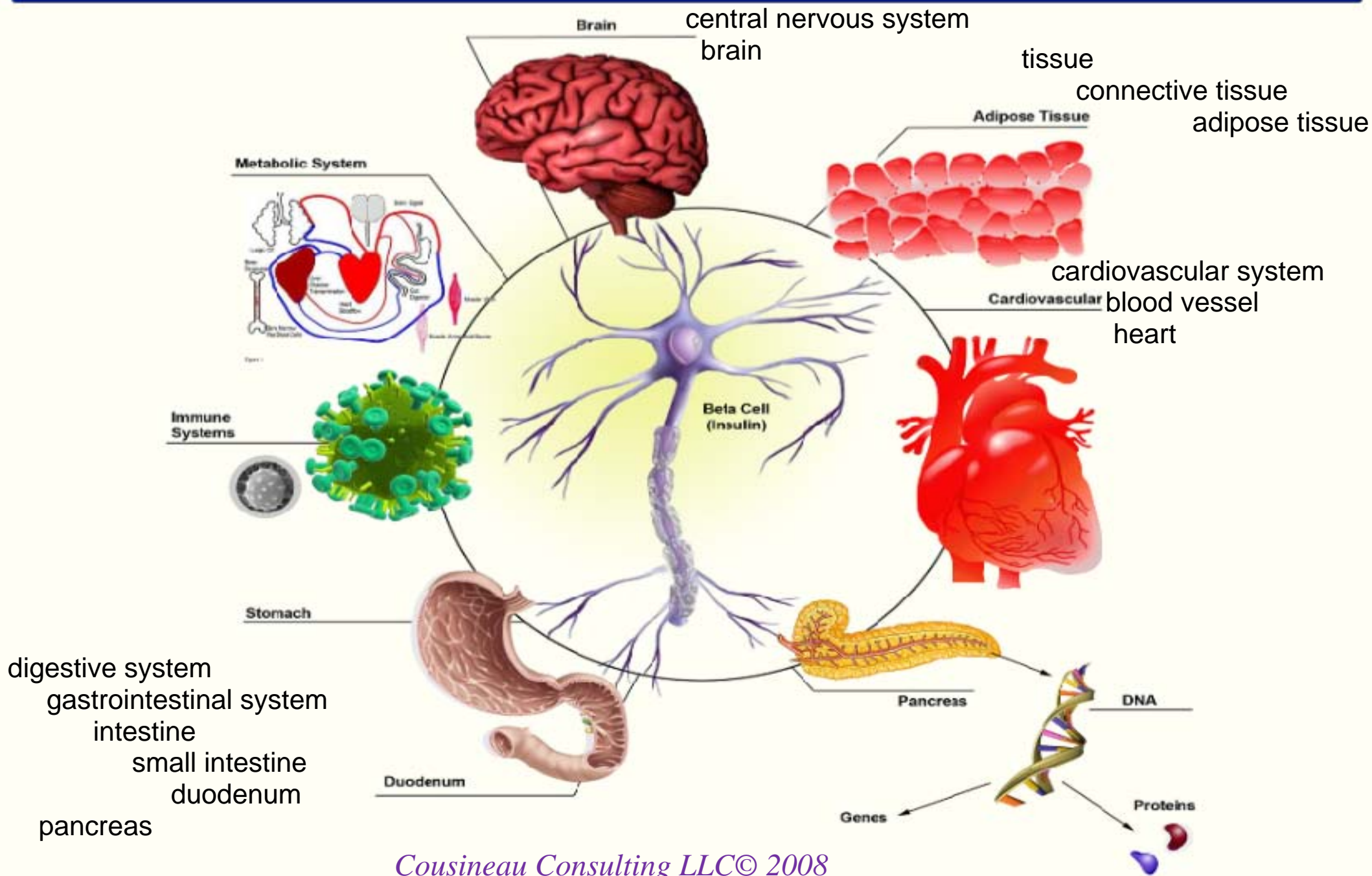


*Genetics*

# Medical Ontology : Relationships between diseases, disorders, & systems, organs and tissues



# Biomedical Ontology : Neuronal interaction between diseases, systems, organs, substances, tissues, cells, proteins and genetics





# Acknowledgments

- Rex Chisholm
- John Osborne
- Julie Zhu & Simon Lin
- NUgene Team - Wendy Wolf, Maureen Smith, Jennifer Allen, Tony Miqueli
- Dong Fu
- UMLS Team
- Medical Knowledge Team – Steve Roessingh, Leo Cosineau, Robert Baud



# Invitation for participation

For Disease Ontology to succeed, it needs buy-in from the community. We hope you will join us in fixing and extending Disease Ontology to meet your needs.