Information Quality for Clinical Knowledge Representation

Kuan-Tsae Huang, Ph.D.
Taskco Corporation
and NLM/NIH
kthuang2@gmail.com

Treasury of e-Health Data

- Help answer big questions
  - Are people who eat less tend to live longer? Why?
  - Any genetic reasons why we have such a high incident rate of cardiovascular disease?
  - How to prevent Kidney stones and treat them?
  - …
- Public Health Intelligence
  - The gathering and analysis of information about health, the causes of ill health, and the patterns and trends of health and ill health in populations
  - Measures to stave off/prevent the onset of disease
  - Prevent drug adverse effects of prescriptions from multiple doctors
  - Success probability of treatment regimen designed for you

OMIM- Online Mendelian Inheritance in Man

NCBI

MIM #610915
Text
Description
Clinical Features
Molecular Genetics
Nomenclature
History
References
Contributors
Creation Date
Edit History

- Clinical Synopsis
- Gene map

Entrez Gene
N Nomenclature
R RefSeq
G GenBank
P Protein
U UniGene

LinkOut

OMIM
Online Mendelian Inheritance in Man
Johns Hopkins University

My NCBI
[Sign In] [Register]

All Databases | PubMed | Nucleotide | Protein | Genome | Structure | PMC | Taxonomy | OMIM

Search OMIM for Go Clear

Limits | Preview/Index | History | Clipboard | Details

Limits: **with Clinical Synopsis**

Display Detailed Show 20 Send to

#610915 GeneTests, Links

OSTEOGENESIS IMPERFECTA, TYPE VIII

Alternative titles; symbols

OI, TYPE VIII
OI8

Gene map locus 1p34

TEXT

A number sign (#) is used with this entry because this form of autosomal recessive osteogenesis imperfecta is caused by mutation in the gene encoding leprecan (LEPRE1; 610339).

DESCRIPTION

Osteogenesis imperfecta is a connective tissue disorder characterized by bone fragility and low bone mass. Due to considerable phenotypic variability, Sillence et al. (1979) developed a classification of OI subtypes based on clinical features and disease severity: OI type I, with blue sclerae (166200); perinatal lethal OI type II, also known as congenital OI (166210); OI type III, a progressively deforming form with normal sclerae (259420); and OI type IV, with normal sclerae (166220). Most forms of OI are autosomal dominant with mutations in one of the 2 genes that code for type I collagen alpha chains, COL1A1 (120150) and COL1A2 (120160). Cabral et al. (2007) described a form of autosomal recessive OI, which they designated OI type VIII, characterized by white sclerae, severe growth deficiency, extreme skeletal undermineralization, and bulbous metaphyses.

CLINICAL FEATURES

Cabral et al. (2007) described 5 patients with a lethal/severe osteogenesis imperfecta-like bone dysplasia caused by mutation in the LEPRE1 gene. The phenotype of the probands overlapped Sillence lethal type II/severe type III osteogenesis imperfecta (see 166210 and 259440), with severe osteoporosis, shortened long bones, and a soft skull with wide open fontanel. However, in contrast to the classic blue sclerae, triangular face, and narrow thorax of severe and lethal osteogenesis imperfecta, their probands had white sclerae, a round face, and a short barrel-shaped chest. Prenatal radiographs demonstrated gracile, undermineralized ribs and long bones. Multiple fractures were present at birth. Long bone radiographs of surviving probands showed bulbous metaphyses and apparent matrix disorganization. Their hands appeared relatively long compared to their forearms, with long phalanges, short metacarpals, and disorganized matrix. Vertebral

Clinical Synopsis

NCBI

MIM #610915
Text
Description
Clinical Features
Molecular Genetics
Nomenclature
History
References
Contributors
Creation Date
Edit History

- Clinical Synopsis
- Gene map

Entrez Gene
N Nomenclature
R RefSeq
G GenBank
P Protein
U UniGene

LinkOut

All Databases | PubMed | Nucleotide | Protein

Search OMIM for

Limits | Preview/Index | History | Clipboard | Details

Limits: **with Clinical Synopsis**

Display Clinical Synopsis Show 20 Send to

#610915

OSTEOGENESIS IMPERFECTA, TYPE VIII

Clinical Synopsis

INHERITANCE :
Autosomal recessive

GROWTH :
Height
Short stature, disproportionate
Dwarfism, short-limbed

HEAD AND NECK :
Head
Wide open anterior fontanelle
Soft skull
Open sutures
Face
Round face
Eyes
White sclerae
Proptosis
Teeth
No dentinogenesis imperfecta

CHEST :
External features
Short, barrel-shaped chest
Ribs, sternum, clavicles, and scapulae
Thin ribs

GENITOURINARY :
Internal genitalia
Inguinal hernia

SKELETAL :
Bone fragility
Severe osteopenia
Normal bone age
Multiple fractures, present at birth
Joint laxity
Skull
Poorly ossified skull
Wormian bones
Spine
Platyspondyly
Scoliosis
Kyphosis
Vertebral compression fractures
Limbs
Thin, gracile long bones
Radial bowing
Femoral bowing
Tibial bowing
Bulbous metaphyses
Externally rotated/abducted legs
Hands
Long phalanges
Short metacarpals

NEUROLOGIC :
Central nervous system
Delayed development

LABORATORY ABNORMALITIES :
Type 1 collagen overmodification
Absent-decreased prolyl 3-hydroxylation at collagen I alpha-1 pro986

MOLECULAR BASIS :
Caused by mutation in the leucine- and proline-enriched proteoglycan 1 gene (LEPRE1, 610339.00

CREATION DATE

Kelly A. Przylepa : 6/11/2007

EDIT HISTORY

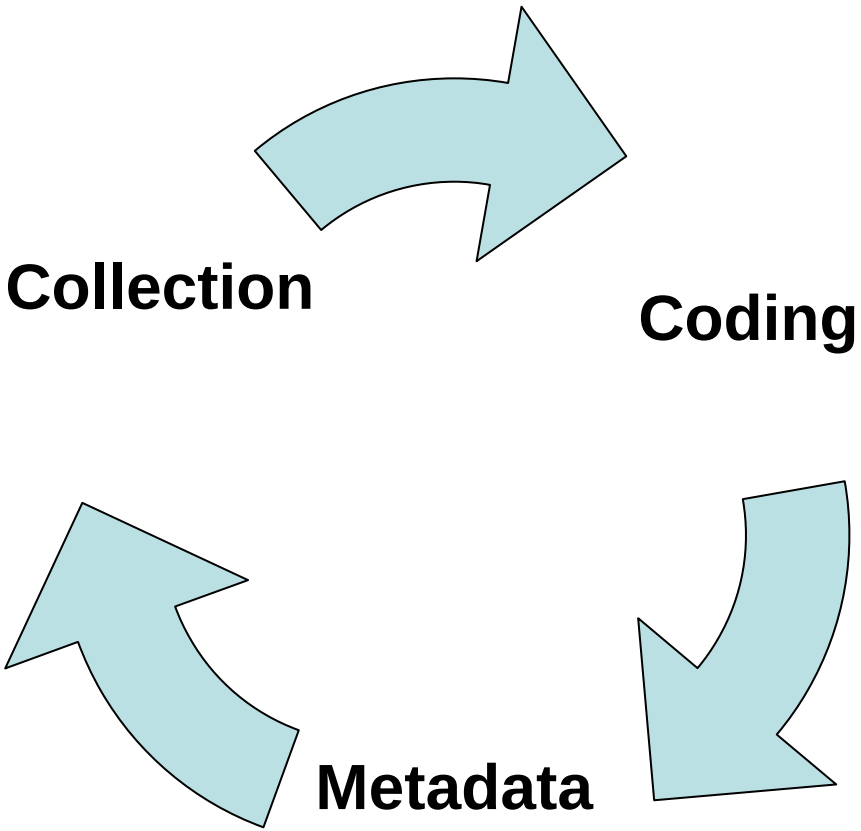
Health Data Nature

- **Diverse data sets –** many and evolving
  - Patient/episode-based
  - >100 data items, including clinical codes
- Varying timetables for submission
- **Distributed –** from 100 - 1000+ magnitude of locations, many IT systems
- **Variety of data providers** – Hospitals, Government, Clinicians, voluntary sector, private sector
- **Share in different contexts –** 1000 magnitude of publications, requests, and questions
- Public Health Statistics
- Multiple ways of collect, coding and interpret of data

Data Sources

- **Different types of data sources**
  - **Public data**
  - **Processed public data**: annotation or indexing
  - **Sensitive data**: individuals or derived from experiments
  - **Special experimental data**: e.g. microarray data
  - Personal research data
  - Team research data
  - **Consortium research data**: group of teams
  - **Personalization data**: individual users
  - **Derived data**: searching/mining of public repositories

Quality Issues



Health Data Collection: A complex task

- From different organizations
  - Collected and produced from centers, clinics, laboratories, etc.
- Heterogeneous
  - Bioinformatics and medical informatics measures, etc.
- Various formats
  - Databases, papers, electronic, XML, etc.
- Various codification rules between organizations
- Data collection form – specify “all” of the research variables of interest
  - As a survey instrument
  - Or a measurement panel
  - Or a questionnaire
- Confidential information (Exclusion)
- Spelling check for medical words
  - **Cytophaga ulginosa** (organism)|**Cytophaga uliginosa** (organism)
  - Infection due to vancomycin resistant **Staphylocccus** aureus (disorder)
  - **Staphylococcus vs Staphylocccus**
  - Glycogenosis viiia
  - N1biii: Extension of tumor beyond the capsule of a lymph node metastasis, < 2 cm in greatest dimension

Impatient/Outpatient Data Sets

- Episode End Date
- Provider Code
- Commissioner Code
- Decided to Admit Date
- Discharge Date
- Date of Birth
- Primary Diagnosis
- First Secondary Diagnosis
- Second Secondary Diagnosis
- Third Secondary Diagnosis
- Primary Operation
- Date of Primary Operation
- Postcode
- Registered GP
- NHS Number (1)
- NHS Number (2)
- Specialty Code
- Administrative Category
- Legal status
- Ethnic Category
- Augmented Care Period 1 Start Date
- Delivery Method
- HRG
- Days in IC and HDU in First Augmented Care Segment
- Admission Method
- Discharge Method
- Consultant Code

- Commissioner Code
- NHS Number (1)
- Postcode
- Registered GP Practice
- Registered GP
- Primary Diagnosis
- Primary Procedure
- Attendance Date
- First Attendance
- Attended or Did Not Attend
- Source of Referral
- Referral Request Received Date

- Patient identifiers
- Research identifiers
- Responsible party identifiers

Need Standardization

- All clinical variables, measurements and survey instruments need to have standard
- Bulk studies – (huge) :
  - Genomic + studies
  - Radiology images
  - more…
- Parameterized versions of #2
- Data collection forms and reusable
  - ACC cardiac cath form(s)
- The same question (set of questions) would be used in many studies and forms

Potential for Errors

- Data entry
  - Unaware of the consequences of inexact or incomplete data on the overall quality of the study
  - Difficult to perform spelling checks on medical/genomic terms
- Samples and questionnaires identification and manipulation
  - Important potential for errors in the processes, numerous manipulations
- Keys generation and management (identification codes)
  - To protect identity and avoid errors in the correspondence between identification numbers and individuals
- Size of databases
  - Millions records
  - Increasing complexity of data transfer, storage, query and analysis
- Validation
  - Essential to insure continuous quality controls, including cross-validations, statistical validations, etc.

Aggregation: Marital Status

Data Set 1	Data Set 2	Data Set 2
Single		Single
Married or Living as Married	Married	Married
Widowed	Widowed	Widowed
Divorced	Divorced	Divorced
Separated	Separated	
	Never Married	
	Living with Partner	
	Refused	
	Don’t know	

Ref: from P3G Observatory project

Questionnaire Form Design

Targeted : Cancers	Ever had cancer	Type of cancer	Onset of symptoms or diagnostic date
Study1	Have you ever had cancer?	What kind of cancer?	In which year was this ascertained?
	Yes, No, I don't know		*Year |_|_|_|_| or age at that time |_|_|*
Study2	Have you ever been told by a doctor or other health professional that you had cancer or a malignancy of any kind?	Has a physician ever told you that you had any of the following cancers?	How old were you when the cancer was first diagnosed?
	Yes, No, Refused, Don't Know	Prostate cancer, Lung or bronchial cancer, Colon or rectal cancer, Bladder cancer, Lymphoma, Other cancer (define)	*|___|___|___| age in years, refused, don't know*
Study3	Has a physician ever told you that you had any of the following cancers?	What kind of cancer was it?	Prostate cancer
	List of cancer		O Never O Before October 2001 O Oct. 2001 - July 2003 O After July 2003

Clinical coding & Medical records

- What fields should be coded?
- In House coding
- Medical coding
- Translations
- Terminologies
- Consistency between studies/sources

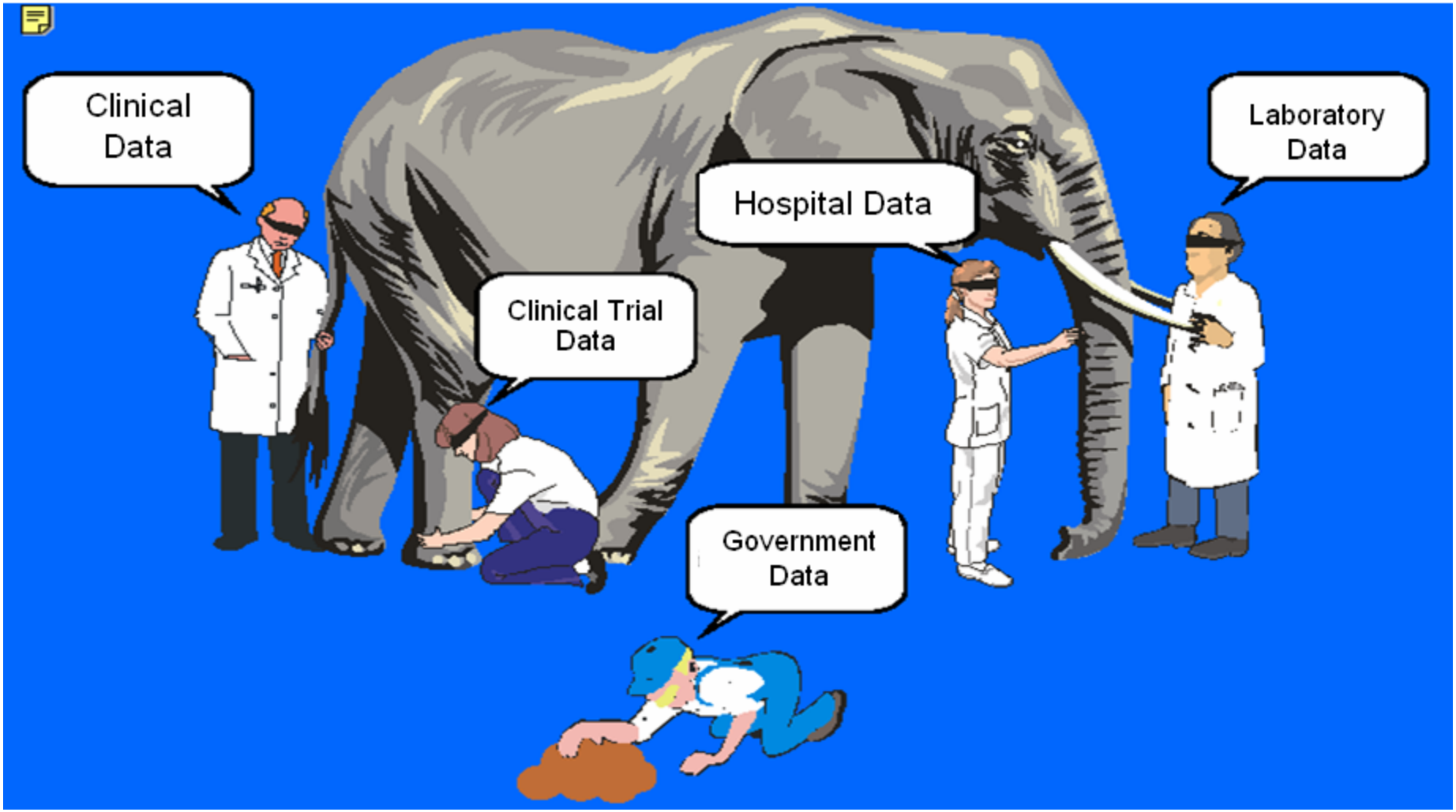
CN# 48555 **Glycogen storage disease, type IX**
CUI C0268147 Concept Status is Reviewed
STY Disease or Syndrome R

Glycogen phosphorylase kinase deficiency (disorder) [SNOMEDCT_2007_01_31/FN
Glycogen phosphorylase kinase deficiency [SNOMEDCT_2007_01_31/PT/
Glycogenosis viiia [SNOMEDCT_2007_01_31/SY
hepatic phosphorylase kinase deficiency [CSP2006/ET
Glycogen phosphoryl kinase def [RCD99/AB
PKD of liver [RCD99/AB/

Metadata

- Common *Models*
  - For data capture, analysis and publication to work together
- Building models requires data and metadata *Services*
  - Protocol designers need access to the latest standards
  - Data collection should be based on the latest terminologies
  - Services need to provide long-lived access to current and versioned data elements (2-5 years for a trial, 10-50 years for follow-up)
- Data Dictionary, including definitions and recording manuals
  - Statistics on quality of data
- Reuse of data in 5 years and beyond
  - Quality and validity of models and data elements
  - Standard data sharing processes

Integration is the key



Context Variation

- Usages: clinical governance, planning, epidemiology, performance management, setting policy
- Stakeholders: Citizen, Public Health Orgs, Parliament, Local Authorities, Researcher, GRO, Researchers, Media, Public, Political parties
- Influences: health policy, devolution, National Statistics, Freedom of Information, data protection, patient involvement, IT developments (eg web), media awareness

Data Aggregation

- Clinical study is a complex task
- Example:
  - Want to know the impact of genes and environment on complex disease
  - Beighton and Versfeld (1985) suggested that type III OI (see 259420) is relatively high in the black population of South Africa
  - By linkage studies, Wallis et al. (1993) excluded the COL1A1 and COL1A2 (120160) loci as the site of the mutation in this form of osteogenesis imperfecta
- Aggregation of data between studies often is needed for a population-based study
- Leverage statistical power for investigation

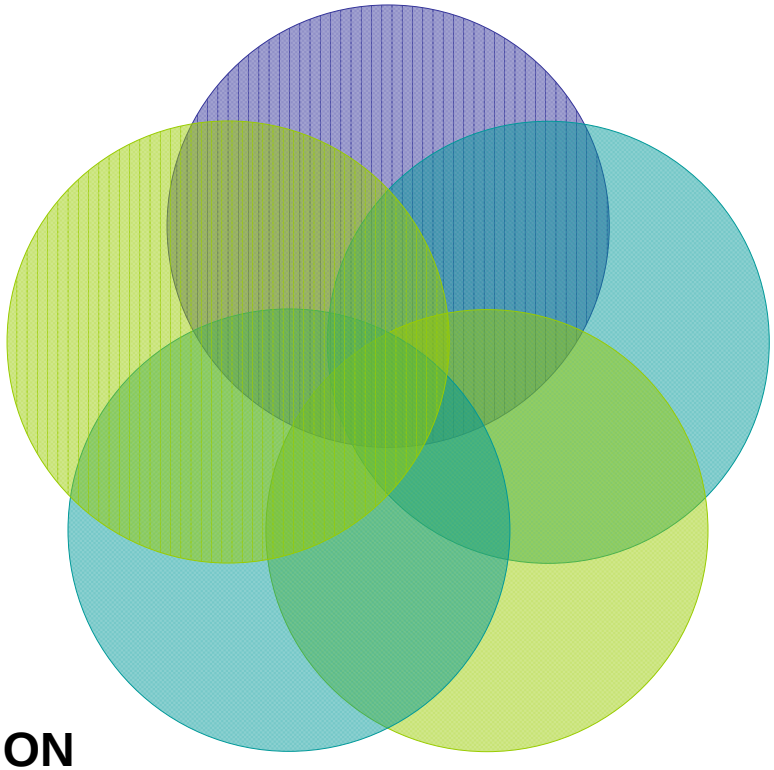
Data Quality Measures

- Validity
- Accuracy
- Completeness
- Fitness for purpose
- Relevance
- Coherence
- Comparability
- Data ‘sign off’

Key to Data Quality

DESIGN OF STUDIES

ETHICS AND GOVERNANCE



INFORMATION COLLECTION/ TREATMENT

INFORMATION HARMOGENIZATION

DATA ANALYSIS

Data Entry Training

Taskco e-Business Corp

Tools for data sharing

- Common *Models with Metadata Services*
  - Description of targeted studies, methods, data, ethics and governance rules, operation procedures, etc
- Spelling Check Tool
  - Difficult to capture errors of medical terms
- Comparison Tool
  - Among the information collected or produced and of procedures used
- Homogenization Tool
  - Schema, distribution formats,
- Knowledge Repository
  - Standard operation procedures or good practices guides
  - Methodologies in epidemiology or genomics

Thank You!!!