

# The Role of Common Data Elements in Artificial Intelligence

*June 4, 2025*

*Denise Warzel, BBA, MSc*

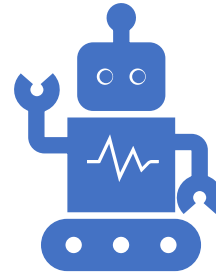
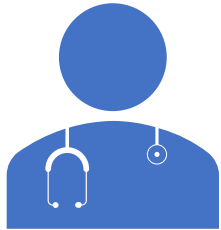
*Semantic Infrastructure Program, Metadata and Models*

*Clinical & Translational Research Informatics Branch (CTRIB)*

*National Cancer Institute (NCI), NIH*

# What if ...


Doctors ***confidently*** use **AI** to specialize treatment for each patient?




***Your Common Data Elements (CDEs) were key to enabling this capability?***

# AI Potential in Biomedical Research ...

**\* Drug Discovery** – Faster access to life-saving treatments (15+ → 5-7 years); Identify new targets; Better prediction of efficacy and safety before expensive trials start; New uses for approved drugs



**Precision Medicine and Personalized Treatments** – Therapies Individualized based on comprehensive patient profiles Including genetic, environmental, lifestyle and medical history; Better predict outcomes and reduce risk at individual patient level

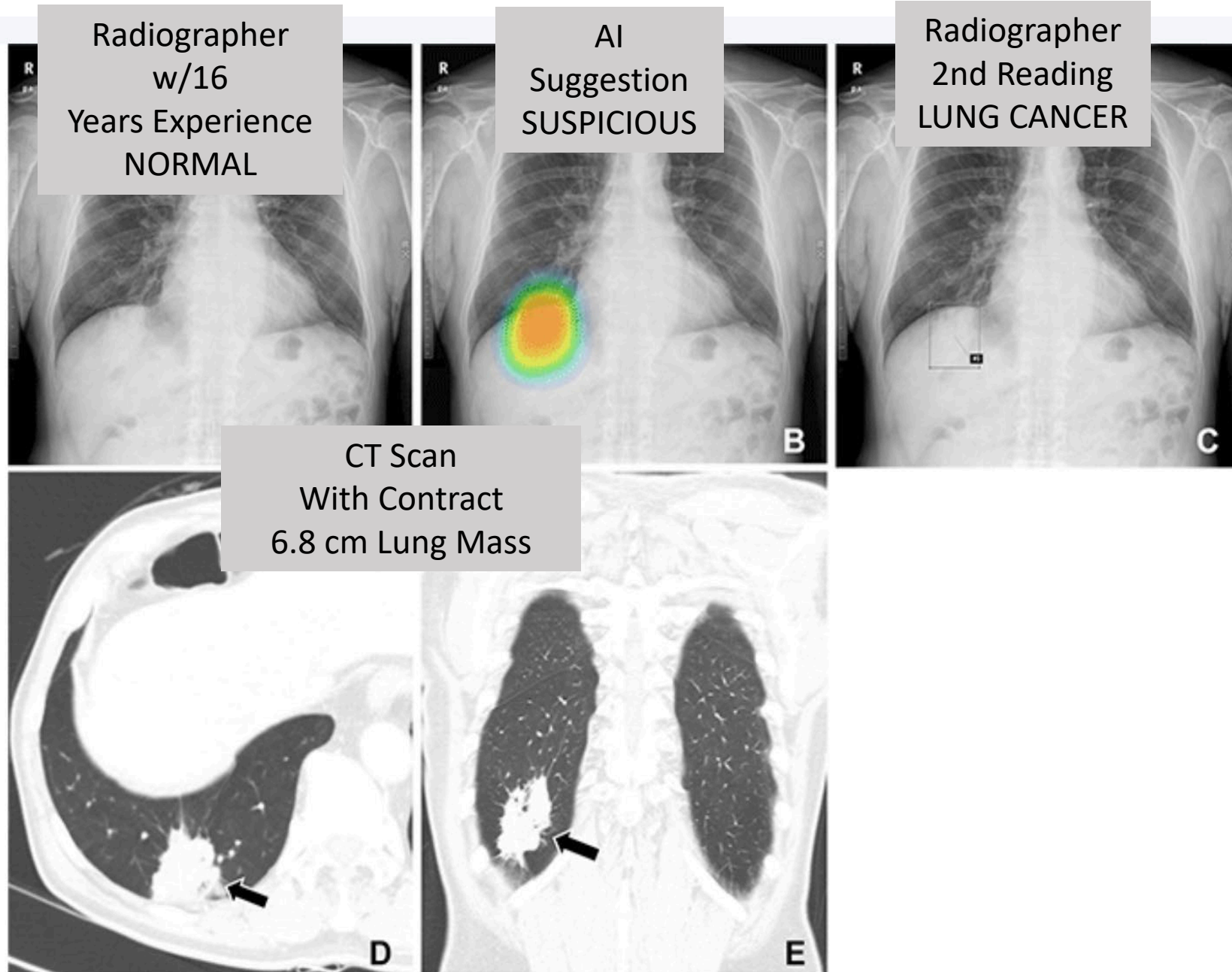


**\*\* Medical Imaging and Diagnostics** – “Superhuman” accuracy and speed leading to earlier detection; Automate Pathology image interpretation: Accuracy, consistency, speed and scalable to areas where experts are scarce

- Jaskaran Preet Singh Saini AI-driven innovations in pharmaceuticals: optimizing drug discovery and industry operations, <https://doi.org/10.1039/D4PM00323C> and Alberto Ocana, A, et al, Integrating artificial intelligence in drug discovery and early drug development: a transformative approach <https://doi.org/10.1186/s40364-025-00758-2>

\*\* Chang Min Park, MD, PhD *Radiological Society of North America* <https://doi.org/10.1148/radiol.222976> ©RSNA 2023

# High Accuracy AI Improves Lung Cancer Detection



# Categories of AI

## Hybrid AI

### Symbolic AI

- Uses logic and rules, leveraging ontologies, knowledge graphs, rule-based inference e.g. rule based expert system for medical diagnostics

### Statistical/Machine Learning (includes LLMs)

- Supervised, unsupervised, reinforcement learning, e.g. Classification systems, Spam detection, Decision trees, hierarchical clustering, self-driving cars

### Neural Network/Deep Learning

- Analyzes and Interprets speech, visual input, audio e.g. Facial recognition, medical imaging, GPT-4 with vision, DALL-E

# Data Quality and Availability

## Garbage In, Garbage Out

- **Manual data cleaning** - time consuming and resource intensive – 70-80% of AI project time
  - Incorrect labels, inconsistent terminology usage, missing data validation
  - e.g. Gestational Diabetes vs Diabetes in Pregnancy
  - Reduces machine readability and **impairs model performance**
  - **Insufficient high quality** training data produces poor quality models

## Format and Structure

- **Inconsistent** data formats
- Heterogenous data models
  - e.g. e.g. “blood pressure” vs “Systolic + Diastolic”, Observation vs Personal Characteristic
- **Lack of Interoperability** reduces reuse in AI pipelines

**Cannot be overcome by algorithms or compute power**

# LLM Hallucinations

## What are “Hallucinations”?

- Factually incorrect information
- Made-up Details
- Wrong information presented “confidently”

## Why does it happen?

- Sophisticated pattern matching
  - Predicts what should come next based on statistical patterns
- Insufficient or poor quality data can lead to incorrect responses

## How to prevent it?

- Use higher-quality data for AI training
- Provide verified datasets to augment the LLM
  - Retrieval-Augmented Generation (RAG)
  - Vector Databases with Embedding
  - Knowledge graphs

**“Guardrails”**



# What is a Common Data Element (CDE)?

---

Metadata

Information  
describe the  
**meaning and  
format of data**



# What is unique about CDEs?

*Deeper Characteristics and Benefits:*

1. **Standard Terminology Concepts** → unambiguous, shared, and computable meaning
2. **Standardized Structure** → machine computability
3. **Independent Semantics** → reusable across physical data models, forms, datasets for interoperability
4. **Persistent Unique Identifier** → identifiable, outside specific data collection systems
5. **Supports FAIR data** → rich metadata, web accessible repository (Findable, Accessible, Interoperable, Reusable)

**Facilitate high performing AI models**



*Interoperability Principle:*  
Shared semantic alignment  
and mapping

# What Do You Mean?

- Context is important in conveying meaning
  - Words have different meanings depending on words around it.
- Some examples:
  - **Agent:** chemical compound or government employee?
  - **Alcohol:** disinfectant or a drink?
  - **Colon:** sentence punctuation or biological organ?
  - **Mole:** animal, blemish, unit of measure, or spy?
  - **Probe:** examination, investigation, or instrument?

→ The above words are **SEMANTICALLY AMBIGUOUS**.
- Words can mean different things in different contexts.
- CDEs linked to standard concepts codes provide ***domain specific context***

# About Ontologies

## → Domain Specific Knowledge Expansion

- “TP53 Gene” Code C17359
- Concept Relationships
  - Gene\_Plays\_Role\_In\_Process
  - Gene\_Associated\_With\_Disease
  - Gene\_Involved\_In\_Pathogenesis\_Of\_Disease
  - Gene\_Has\_Abnormality
  - Gene\_Found\_In\_Organism
- Knowledge represented through Concept Relationships
  - Enhances LLM’s
  - Provides Semantic Context
  - Helps Reduce Hallucinations

Role Relationships ( 49 ) <a href="#">[top]</a> asserted or inherited, pointing from the current concept to other concepts:		
Relationship <a href="#">↑</a>	Related Code <a href="#">↑↓</a>	Related Name <a href="#">↑↓</a>
Gene_Associated_With_Disease	C9325	Adrenal Cortical Carcinoma
Gene_Associated_With_Disease	C4872	Breast Carcinoma
Gene_Associated_With_Disease	C3099	Hepatocellular Carcinoma
Gene_Associated_With_Disease	C3871	Nasopharyngeal Carcinoma
Gene_Associated_With_Disease	C9145	Osteosarcoma
Gene_Associated_With_Disease	C4815	Thyroid Gland Carcinoma
Gene_Found_In_Organism	C14225	Human
Gene_Involved_In_Pathogenesis_Of_Disease	C187447	A53 Diffuse Large B-Cell Lymphoma
Gene_Involved_In_Pathogenesis_Of_Disease	C9094	Adult Glioblastoma
Gene_Involved_In_Pathogenesis_Of_Disease	C6650	Ampulla of Vater Adenocarcinoma
Gene_Involved_In_Pathogenesis_Of_Disease	C9477	Anaplastic Astrocytoma
Gene_Involved_In_Pathogenesis_Of_Disease	C8374	Bowenoid Papulosis
Gene_Involved_In_Pathogenesis_Of_Disease	C9119	Breast Medullary Carcinoma
Gene_Involved_In_Pathogenesis_Of_Disease	C9039	Cervical Carcinoma
Gene_Involved_In_Pathogenesis_Of_Disease	C5136	Childhood Glioblastoma
Gene_Involved_In_Pathogenesis_Of_Disease	C2955	Colorectal Carcinoma
Gene_Involved_In_Pathogenesis_Of_Disease	C4025	Esophageal Adenocarcinoma
Gene_Involved_In_Pathogenesis_Of_Disease	C4024	Esophageal Squamous Cell Carcinoma
Gene_Involved_In_Pathogenesis_Of_Disease	C9166	Gallbladder Adenocarcinoma
Gene_Involved_In_Pathogenesis_Of_Disease	C4052	Vulvar Squamous Cell Carcinoma
Gene_Plays_Role_In_Process	C16269	Aging
Gene_Plays_Role_In_Process	C16397	Cell Cycle Process
Gene_Plays_Role_In_Process	C19598	Cell Cycle Regulation Process
Gene_Plays_Role_In_Process	C16513	DNA Repair
Gene_Plays_Role_In_Process	C20150	Positive Regulation of Apoptosis
Gene_Plays_Role_In_Process	C19077	Transcriptional Regulation
Gene_Plays_Role_In_Process	C26040	Tumor Suppression

# Characteristics of Well-Curated CDEs

## Machine Readable

- Computers can easily process and interpret the meaning of data

## Standardized

- Consistent structure supports reuse for data validation, semi-automated mapping and transformation

## Semantic Clarity

- Standard concepts eliminate ambiguity, ensure consistent understanding



# Key messages about CDE Concepts

01

Words can have  
different meanings

02

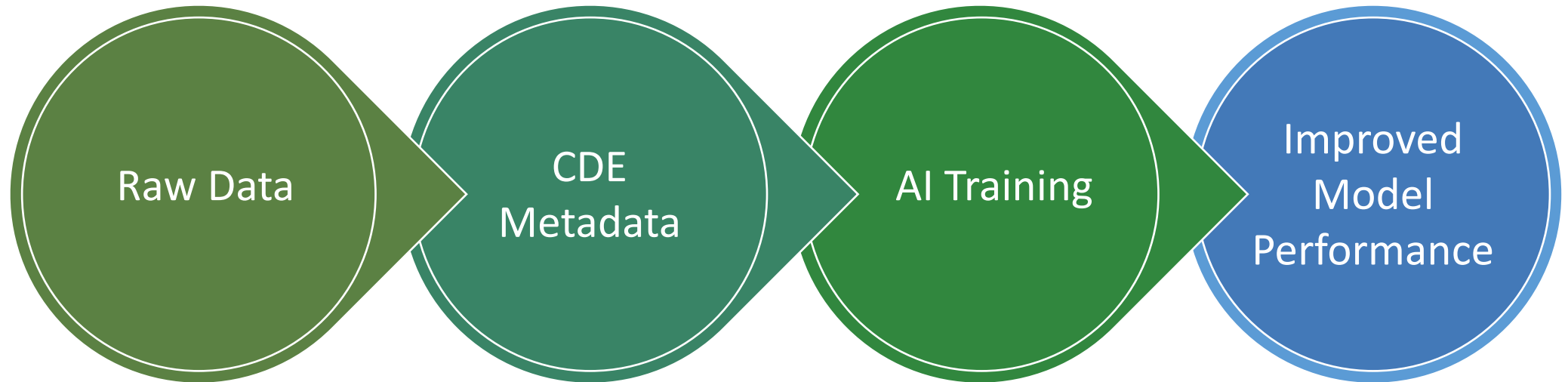
CDEs use standard  
concepts codes  
providing  
unambiguous, shared  
meaning

03

Concepts provide  
domain specific  
context, linked to  
knowledge, expanding  
human and machine  
understanding



# Enhancing AI Models with CDEs



# CDEs Help Address AI challenges

## Improve Data Quality

- Consistent, computable semantics and labeling
- Structured data collection and validation

*Unambiguous data collection methods reduces uncertainty improves supervised learning accuracy*

## Increase Data Availability

- Common data formats increase interoperability and pooling of datasets
- Rich CDE metadata facilitates automated mapping reducing labor intensive data preparation across different sources

*More high quality data with less effort speeds up AI-pipeline for AI Model development and training*

## Reduce LLM Hallucinations

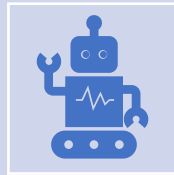
- Improve retrieval quality
- Provide semantic grounding
- Enable ontology and knowledge graph integration

*Contextualized CDE-metadata based training and standardized data support more factual results*

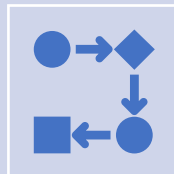
# Future Directions



**Expand CDE Usage**



**Integrate with Emerging AI  
Technologies**



**Continuous Improvement**

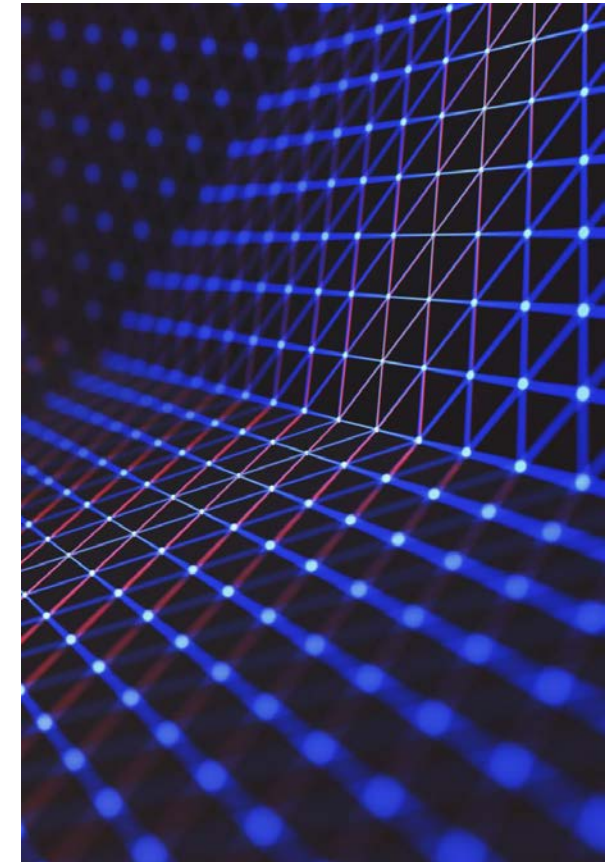


## *Call to Action:*

Adopt CDEs to Help Make Data AI Ready

→ Advance Research and Improve Healthcare

Improve	Support	Enhance
Improve data semantics and consistency	Support harmonization, mapping and transformation	Enhance knowledge acquisition to accelerate discoveries
Data Quality	Data Availability	Data analytics



Thank you!

A thick, hand-drawn orange line that spans the width of the text above it, positioned below the 'Thank you!' message.