CM

- Dr. Tavpritesh Sethi
- This objective of this course is to train students in creation of computing solutions that are relevant to medicine. The course will pick timely and relevant topics in computing such as the building blocks of electronic health records, modeling and visualizing diseases, pathogen and human factors involved in spread and the ever increasing role of information management and computing in managing diseases.
- Dear Students, Quiz 1 will be held on Thursday, 28th August. The syllabus will include all topics covered up to the Tuesday class. The paper will be a combination of subjective and objective questions. Please make sure to carry your ID card.
- TAs
    - Students can reach out to the following TAs during their support hours in Room A316. Please make use of these slots for doubt clarification.

```
TA Name Day Time slot   Gmail
Pallawi Tuesday 2-3pm   pallawik@iiitd.ac.in
Mansi   Monday  4-5pm   mansigo@iiitd.ac.in
Varsha  Thursday    2-3pm   varshar@iiitd.ac.in
Abuzar  Monday  4-5pm   abuzark@iiitd.ac.in
Prateek Tuesday 2-3pm   prateek24307@iiitd.ac.in
Vibhuti Monday  4-5pm   vibhutik@iiitd.ac.in
Abhinav Tuesday 2-3pm   abhinavs@iiitd.ac.in
```

Table of Contents

Computing for Medicine, Lecture 1: Introduction

(For Monsoon 2025, as per C4M-Lecture-1_Introduction.pptx_compressed.pdf)

---

1. **Course Overview**

- **Aim:** Learn how computing and data science apply to medicine and healthcare.
- **Key Focus:** Interoperable health data, data science, machine learning (ML), AI, real-world health datasets.
- **Skill Development:**
  - Build systems for safe medical data exchange.
  - Critique and apply modeling for health data.
  - Design solutions for actual healthcare problems.
  - Learn inclusion, safety, and ethics in healthcare AI.

---

2. **Course Structure**

- **The Big Picture:** How computing interacts with healthcare systems.
- **Data:**
  - Spotlight on *FAIR* (Findable, Accessible, Interoperable, Reusable) and *Open* data.
  - Standards:
    - *Syntactic Interoperability:* HL7, DICOM, FHIR (for data exchange).
    - *Semantic Interoperability:* SNOMED, Ontologies (for meaning).
- **Data Science, ML & AI:**
  - *Structured data:* Statistics, basic ML.

-
    - *Unstructured data:* NLP (Natural Language Processing) in healthcare.
    - *New Topic:* Agentic AI—AI that takes action itself in healthcare.
- **Case Studies:**
    - Exploring/researching open health datasets.
    - Real examples: Antibiotic resistance, ICU predictive modeling.
    - Examining inclusion, safety, ethics.

---

3. **Grading & Assignments**

Component	Weight
Assignments	15%
Quizzes	15%
Mid Sem Exam	20%
End Sem Exam	25%
Project	25%

Quizzes: “n-1 policy” (miss one without penalty).

Projects and *activities* are central—learn by doing.

---

4. **Academic Integrity**

- *What counts as plagiarism:*
    - Copying homework.
    - Using someone else’s work/ideas without citation.
- *Consequences:*
    - First time: Zero marks.
    - Second time: Reported to committee.
    - If in the final report: Grade goes down by one.

---

5. **Biological Systems Are Unique**

- **Features:**
    - Adaptive, self-organizing, constantly changing, oscillatory, far from equilibrium, sensitive, not always optimal.
- *Physical efficiency* (like perfect networks in lungs) can actually make a system more fragile—not always optimal for health (example: bronchial tree and asthma).

---

6. **Interdisciplinary Nature**

- This field combines computer science, biology, medicine, and social science.
- Inspired by coursework/programs at Stanford, Harvard, Melbourne, and others.

---

7. **Uses of Health Data**

- **Applications:**
  - Individual care (diagnosis/treatment)
  - Public health planning
  - Policy evaluation
  - Improving safety, cost-effectiveness, outcomes
- *Same data* can answer different questions—depends on how/why it is used.

---

8. **Types of Medical Studies**

- **Retrospective:** Look back at old records.
- **Prospective:** Start now, collect data moving forward.

---

9. **Why Data Science Is Important in Health**

- Helps:
  - Identify risk factors (e.g., cholesterol → heart disease).
  - Control for "confounders", *hidden factors* that can mislead results.
- *Simpson's Paradox:* Trends in subgroups may be opposite to the whole population. Need care in data interpretation.

---

10. **Types of Data Analysis Problems**

- Must think about *causality*: What causes what?
- *Forks (confounding):* Multiple factors affect a result.
- *Colliders:* Sometimes including too many variables hides real effects.

---

11. **AI Failures**

- **Common failures:**

  - Biased data (e.g., only light skin in melanoma training data).
  - Human error.
  - Regulatory issues.
  - Over-trusting complex, "black box" models.
  - Disinformation via AI tools.

- *Case Study:* An AI learned that pneumonia patients with asthma had better survival, but only because they got better care—not because asthma was protective.

---

12. **Making Data FAIR**

Principle	What it means
Findable	Data has a persistent unique identifier, rich metadata

Principle	What it means
Accessible	Data can be retrieved with standard protocols, metadata always available
Interoperable	Data uses shared vocabularies and formats
Reusable	Data has clear licenses, provenance, and meets community standards

---

13. **Major Example Health Datasets**

- *CORD-19*: COVID-19 scholarly articles dataset.
- *MIMIC-III*: ICU patient data (demographics, labs, notes).
- *Johns Hopkins COVID* datasets: For specific clinical research.
- *Vivli*: Clinical trials and antimicrobial resistance data.

Most are open to global researchers for free and used for real-world experiments and studies.

---

14. **Key Takeaways**

- Computing can transform how medicine is practiced, but ethical, safe, and reproducible ways of handling data are critical.
- Interdisciplinary skills are necessary: programming, statistics, domain knowledge.
- Real-world datasets are available to explore and build medical AI and tools.
- Always consider data limitations and risk of bias.
- Plagiarism carries heavy consequences—always cite and do your own work.

Computing for Medicine, Lecture 2: The Big Picture (Continued)

(Based on C4M-Lecture-2_Big_Picture_Contd.pptx_compressed.pdf)

---

1. **Case Study: Simpson's Paradox in Vaccine Data**

- **Data Example:** COVID-19/Israeli Data (Aug 2021)
  - **Age Groups:** All ages, Under 50, Over 50
  - **Vaccine Efficacy:**
    - Efficacy appears high in each age group for vaccinated people.
    - But "overall" efficacy looks much lower.
  - **Cause:** Older people (at higher risk) are more vaccinated.
- **Concept:**
  - *Simpson's Paradox* — When trends are present in separate groups but disappear or reverse when groups are combined.
  - **Moral:** Always stratify data and check subgroups before drawing conclusions.

---

2. **Digitizing Human Health: Challenges**

- **Real World:** Health is *continuous* (analogue signals), e.g., heart rate, blood pressure.
- **Computing World:** Works with *digital* (discrete numbers, finite precision).

- **Implication:** Computational models always approximate real health data—never fully capture all details.

---

3. **Types of Medical Studies: Recap**

- **Retrospective Study:** Looks back at existing records.
  Example: Checking past patient histories to find links between exposure and disease.
- **Prospective Study:** Follows participants forward in time and records events as they happen.
  Example: Recruit healthy people, track who develops disease over the years.

---

4. **Conditional Probability and Confounding**

- **Conditional Probability:** Probability that one event will occur given that another HAS happened (used to analyze risk factors).
- **Confounding:** A hidden variable affects both the cause and the effect, making it seem like there is a direct link.
  - *Example:* Ice cream sales and drowning—both go up in summer, but summer is the confounder.

---

5. **Simpson's Paradox in Depth**

- **Description:**
  - Analysis on subgroups/different slices of data can show the opposite trend to analysis on full data.
- **Practical Advice:**
  - Always check data both with and without grouping by possible confounders (e.g., age, risk categories).

---

6. **AI in Healthcare: Potential and Pitfalls**

Types of Failures

- **Biased Data:** Training data does not represent all populations (e.g., an AI for skin cancer trained only on light skin fails on dark skin).
- **Epistemic Failure:** Model makes confident predictions on things it was never trained (e.g., confusing Covid for a cat).
- **Regulatory Failure:** No clear rules for safe deployment, leading to risk and uncertainty.
- **Human Error:** Misinterpreting the output, overtrusting or misunderstanding limitations.
- **Disinformation:** AI tools can be misused to produce fake, misleading or harmful medical information at speed.

Case Studies

- *Pneumonia/Asthma Example* (Caruana et al. 2015):
  - A model wrongly concluded asthma patients with pneumonia have *lower risk*.

  - Why? These patients got much more intensive care, so the data reflected better outcomes, not a lower biological risk.
- *Dedicated AI vs. LLMs for Diagnosis (2025 study):*
  - Dedicated medical expert systems (built for years) still slightly outperform general-purpose LLMs (like ChatGPT, Gemini) for clinical case diagnosis, especially without lab results.
  - When labs are included, differences narrow.

The “Halo Effect”:

- People often over-trust complex models just because they are “deep” or “new”—not always justified.
- Transparent, interpretable models are essential in medicine, even if their accuracy is a bit lower.

---

7. **Disinformation and the “Misinfodemic”**

- **Infodemia:** Rapid spread of too much info—much of it incorrect/dangerous.
- **Recent Observations:**
  - Researchers made 102 fake medical blog posts with lots of disinfo using an LLM in about an hour.
  - Posts targeted specific groups and used fake testimonials and references.
  - Images were fabricated in minutes.
- **Lesson:**
  - AI speeds up the spread of fake health information (“weapons of mass disinformation”).
  - We need transparency, human vigilance, regulatory oversight, and robust “guardrails” to prevent harm.

---

8. **Open Discussion: Who Benefits from Health Data?**

- **Beneficiaries:**
  - Patients (better diagnosis/care), policymakers, researchers, tech companies, public health systems.
- **Risks:**
  - Misuse of sensitive health info, spread of harmful fake news, bias against vulnerable groups.

---

9. **Key Takeaways**

- **Always consider data context and structure before trusting analytic results.**
- **AI in medicine needs interdisciplinary knowledge, careful design, and oversight.**
- **Transparent, fair, and ethical use of health data benefits everyone—but misuse can be dangerous.**
- **Understanding bias, confounding, and paradoxes like Simpson’s Paradox is crucial for trustworthy medical computing.**

Computing for Medicine, Lecture 3: Data Sources

(Based on C4M-Lecture-3_Data_Sources_compressed.pdf)

---

1. **Who Benefits from Health Data?**

- Patients receive better diagnosis and treatment.
- Healthcare providers improve care quality and coordination.
- Researchers advance medical knowledge.
- Policymakers design effective health interventions.
- The public gains from better health systems and disease control.

---

2. **Starting Point: Healthcare Data**

- Healthcare data often starts with **patient records** from visits.
- Example of a medical record:
  - Patient symptoms: cough, fever, dyspnea.
  - Physical exam results: blood pressure, pulse, fever measurement.
  - Lab data: blood tests, fecal occult blood.
  - Imaging: chest X-rays.
  - Medications prescribed and dosage.
  - Follow-up visits with changes in symptoms and tests.
- These records have multiple problems tracked over time, showing progression and treatment response.

---

3. **FAIR Principles for Medical Data**

- To **Make Healthcare Data Reproducible and FAIR**:
  - **Findable:** Assign **unique, permanent IDs** and rich metadata.
  - **Accessible:** Data must be retrievable with open, standard methods; even if data is removed, metadata stays available.
  - **Interoperable:** Use common languages/vocabularies that all systems can understand and link to other data.
  - **Reusable:** Clear licenses, provenance (history of data), and community standards ensure data can be safely reused.

---

4. **Why Use Standards?**

- Standards allow the **same medical vocabulary** to be shared across systems.
- For example, coding systems like:
  - **ICD-10:** International Classification of Diseases for diagnoses.
  - **SNOMED CT:** Detailed clinical terms.
  - **LOINC:** Laboratory test codes.
  - **RxNorm:** Drug prescriptions.
  - **HL7, FHIR:** Standards for exchanging data electronically.
- With standards, computers can **understand and process** medical records accurately.

---

5. **How Does a Computer Read a Medical Record?**

- Medical records are converted into **standardized codes** and data formats.
- Example of one visit:

  - Patient characteristics encoded: 60 years old, male, obese.
  - Symptoms and tests coded with SNOMED CT or LOINC.
  - Diagnosis coded with ICD-10.
  - Medication coded with RxNorm.
- This enables computers to:
  - Store the data.
  - Exchange it with other systems.
  - Use it for analysis or decision support.

---

6. **Structured vs. Unstructured Data**

Type	**Description**	**Example**
Structured Data	Organized in clear fields, tables	ICD codes, prescriptions, vitals
Unstructured Data	Free text or reports	Doctor's notes, discharge summaries, radiology reports

- Both types are important; structured data is easier for computers, but unstructured data contains rich clinical details.

---

7. **What is an Electronic Health Record (EHR)?**

- *ISO Definition:* An EHR is a securely stored, computer-processable collection of a patient's health info.
- It supports ongoing healthcare and includes data from past, current, and future care.
- Accessible by authorized providers to improve quality and coordination.
- Contains diverse data: consultations, tests, prescriptions, referrals, hospital stays, and more.

---

8. **Purpose of EHRs Across Countries**

- Examples:
  - **Australia’s HealthConnect:** Focus on patient care, quality, and management.
  - **Austria’s ELGA system:** Includes patient care and financial/administrative workflows.
- Common goals:
  - Improve patient care.
  - Help doctors make informed decisions.
  - Facilitate research.
  - Reduce errors and streamline workflows.

---

9. **Interoperability in Healthcare**

- **Definition (IEEE 1990):** Ability of systems to exchange and use information.
- **Levels of interoperability:**
  - *Technical:* Ability to send/receive data.
  - *Semantic:* Recipient understands the data.

  - *Process:* Data is used effectively in workflows.
  - *Human/Clinical:* Effective use improves patient care.
- Interoperability **enables AI, big data**, better medical communication, research, and global cooperation.

---

10. **Examples of Interoperability Standards**

- **HL7 ADT messages** (for patient admissions, transfers, discharges).
- **FHIR Standard** uses JSON format to exchange medical data in APIs.
- These allow efficient, standardized data exchange across healthcare IT systems.

---

11. **Key Open Health Datasets**

- **COVID-19 Open Research Dataset (CORD-19):** Scholarly articles on COVID-19 and coronaviruses.
- **MIMIC:** Intensive care unit patient data.
- **Johns Hopkins COVID-19 Collaborative:** Patient records from Epic EHR updated weekly.
- **National COVID Cohort Collaborative (N3C):** Combined patient-level COVID data from many institutions.
- Other datasets from Zenodo, Figshare, GitHub, Harvard Dataverse, ImmPort, and more.

---

12. **Special Mention: Antimicrobial Resistance Data**

- Example software: AMRsteward AI Dashboard for antibiotic stewardship.
- Open-source, designed to track and optimize antibiotic use in hospitals.
- Patient data used to monitor infection and resistance trends.

---

13. **Summary**

- **Data is the foundation of modern medicine's computing revolution.**
- Standards and interoperability enable **safe, effective sharing and analysis**.
- Combining structured and unstructured data gives a complete health picture.
- FAIR principles ensure data can be reused reliably and ethically.
- Access to rich, open datasets accelerates research and innovation.
- Understanding healthcare data sources is key to making impactful AI and computational health tools.

Thanks for reading!

Computing for Medicine, Lecture 4: Semantic Interoperability

(Based on C4M-Lecture-4_Semantic_Interoperability_compressed.pdf)

---

1. **What is Interoperability?**

- **Definition:** The ability of different systems or components to *exchange information* and *use* the exchanged information effectively.

- **Levels of Interoperability:**
    - **Technical:** Systems can send/receive data.
    - **Semantic:** Systems understand the meaning of data received.
    - **Process:** Data is used properly in workflows or decisions.
    - **Human (Clinical):** Data improves patient care in meaningful ways.

Semantic interoperability is key to making digital medicine effective by ensuring data is not just exchanged but also understood.

---

2. **Why Semantic Interoperability Matters**

- Many medical data sets today are isolated in incompatible systems.
- Lack of interoperability slows medical progress.
- Technologies like AI, big data analytics, and mobile apps depend on interoperable, meaningful data.
- Broad domains impacted:
    - Artificial intelligence and big data.
    - Medical communication among caregivers.
    - Research that combines data across institutions.
    - International cooperation on diseases and treatments.

---

3. **Sharing Patient Data to Improve Care**

- Studies show *electronic health information exchange* among hospitals improves coordination.
- Patients benefit as providers share data quickly, reducing repeated tests and errors.
- Hospitals active in Health Information Organizations (HIOs) more often share patient care successfully.

---

4. **Semantic Interoperability with Text**

- Medical data often comes as *unstructured text* (e.g., clinical notes).
- To make use of such data:
    - Extract meaningful terms.
    - Turn them into standardized forms computers can analyze.
- Example clinical note about allergies includes symptoms, medication history, and doctor assessments in narrative form.
- Semantic tools help computers interpret this kind of textual, complex data.

---

5. **Terminologies and Ontologies: Tools for Semantic Interoperability**

Term	What It Means
Terminology	Standardized set of terms to describe a domain (e.g. medicine).
Thesaurus	Groups similar or related words to help with searching.
Controlled Vocabulary	A list of preferred terms that limits variation for consistency.

Term	What It Means
Classification	Organizes terms into mutually exclusive groups (e.g., ICD codes).
Ontology	Defines concepts and relationships formally for computers to reason about.

- Ontologies are the most powerful, representing concepts, attributes, and relationships logically to help machines deeply understand data.

---

6. **Common Clinical Ontologies and Terminologies**

Use Case	Ontology/Standard	Examples
Diagnoses	SNOMED CT, ICD, Orphanet, NCIT	Breast carcinoma, Rare genetic diseases
Phenotypic abnormalities	Human Phenotype Ontology (HPO)	Specific clinical symptoms
Medications	RxNorm, DrugBank, ChEMBL	Panobinostat (a drug)
Adverse drug reactions	Ontology of Adverse Events (OAE)	Injection-site reaction
Procedures	Medical Dictionary for Regulatory Activities (MedDRA)	Cardiac surgery
Lab tests	LOINC	Serum creatinine measurement
Imaging data	DICOM, RadLex	Bone thinning in X-rays

---

7. **What are Ontologies?**

- Ontologies provide both a **shared vocabulary** and the **rules/constraints** for how those terms relate.
- They define:
  - Concepts (things like diseases, symptoms).
  - Attributes (characteristics like severity, location).
  - Relationships (e.g., a disease *affects* a body part).
- Use **description logic** to allow computers to reason about data meaningfully.

---

8. **SNOMED CT: A Key Clinical Ontology**

- Most comprehensive clinical healthcare terminology worldwide.
- Used in over 80 countries for EHR documentation and reporting.
- Includes >300,000 concepts grouped into 10 axes:
  - Topography (body parts)
  - Morphology (cell/tissue changes)
  - Organisms (bacteria, viruses)
  - Chemicals and drugs
  - Signs and symptoms

  - Diagnoses
  - Procedures
  - Social context and jobs
  - Devices and agents
  - General qualifiers
- Concepts have:
  - Unique machine-readable IDs.
  - Human-readable names (Fully Specified Name and synonyms).
  - Relationships to define meaning precisely.
- Allows *pre-coordinated* terms (single codes) or *post-coordinated* expressions (combining codes for more detail).

---

9. **How Ontologies Help Computer Understand Clinical Data**

- Example: Viral pneumonia defined as
  - A type of infective pneumonia.
  - Caused by a virus.
  - Found in the lungs.
- This logical structure helps in advanced querying, inference, and decision support.

---

10. **Clinical Data Integration Examples**

- Johns Hopkins initiatives use such standards and ontologies to:
  - Combine COVID-19 patient data from many hospitals.
  - Create detailed, computable common data models.
  - Help researchers analyze complex medical questions reliably.

---

11. **Resources for Health Data and Research**

- Repositories & data sources include:
  - 4CE Consortium (COVID-19 clinical data).
  - Figshare, GitHub COVID-19 data.
  - NIH Data Repositories.
  - MIMIC-III (ICU patient data).
  - Vivli (global clinical trial data sharing platform).
- Open data accelerates medical research, improves AI models, and supports transparent science.

---

12. **Summary**

- Semantic interoperability makes healthcare data **meaningful and shareable** across systems.
- Controlled vocabularies and ontologies like SNOMED CT allow computers to **understand complex clinical concepts**.
- This understanding powers:
  - Better clinical decision-making.
  - Advanced AI in healthcare.

    - Collaborative research.
- Standards adoption is essential as medicine becomes more data-driven and digital.

Thanks for following along!

Computing for Medicine, Lecture 5: SNOMED CT

Computing for Medicine - Lecture 5

Overview

SNOMED CT (Systematized Nomenclature of Medicine Clinical Terms) is a comprehensive, multilingual clinical healthcare terminology system designed to support the electronic exchange of clinical health information.

---

1. Core Architecture: Concepts and Descriptions

Concepts - The Foundation

- **Definition**: Clinical meanings that remain constant and unchanging
- **Characteristics**:
    - SNOMED CT is fundamentally concept-oriented
    - Each concept has a unique machine-readable Concept ID
    - Concepts represent clinical ideas independent of language or presentation

Descriptions - Human Interface

Purpose: Provide human-readable representations of concepts

Types of Descriptions:

1. Fully Specified Name (FSN)

- **Purpose**: Unique and unambiguous identification
- **Target Audience**: NOT for end users (technical/administrative use)
- **Format**: Contains suffix in parentheses indicating primary hierarchy
- **Example**: `myocardial infarction (disorder)`
- **Importance**: Ensures precise concept identification across systems

2. Display Term

- **Purpose**: User-friendly representation in specific language
- **Format**: Often the FSN without the hierarchical suffix
- **Example**: `myocardial infarction` (derived from FSN above)
- **Usage**: Primary term shown to clinicians and patients

3. Synonyms

- **Purpose**: Alternative ways to express the same concept
- **Classification**:
    - **Preferred**: Recommended alternative term
    - **Acceptable**: Valid but not preferred alternative
- **Benefit**: Accommodates different clinical practices and regional variations

---

2. Relationship Structure

Supertype Relationships (IS-A Hierarchy)

- **Universal Rule**: Every concept (except root) has a supertype
- **Concept ID**: `116680003` for the |is a| relationship itself
- **Function**: Creates taxonomic hierarchy for inheritance and classification
- **Example Hierarchy**:

  ```
  Clinical Finding → Disorder → Cardiovascular Disorder → Myocardial
  Infarction
  ```

Defining Attributes

- **Purpose**: Specify additional characteristics beyond hierarchy
- **Usage**: Combined with supertype to fully define concepts
- **Examples**:
    - Location attributes (e.g., "finding site")
    - Temporal attributes (e.g., "onset")
    - Severity attributes

---

3. Concept Definition Completeness

Sufficiently Defined Concepts

- **Criteria**: Defining relationships are adequate to distinguish from:
    - All supertype concepts
    - All sibling concepts (same level in hierarchy)
- **Implication**: Can be automatically classified by reasoning systems
- **Benefit**: Enables precise automated inference

Primitive Concepts

- **Definition**: NOT fully defined by their relationships
- **Limitation**: Lack unique relationships to distinguish from parent/sibling concepts
- **Example**: "Pneumonia" without specific defining characteristics
- **Current State**: Large portions of SNOMED CT remain primitive
- **Impact**: Requires manual classification and limits automated reasoning

Equivalence Principle

- **Rule**: Concepts with identical defining characteristics are either:
  - Equivalent to each other, OR
  - One is a subtype of the other
- **Importance**: Maintains logical consistency in the terminology

---

4. Clinical Expression Methods

Pre-coordinated Expressions

- **Definition**: Single concept identifier represents complete clinical idea
- **Format**: `ConceptID |Concept Name|`
- **Advantages**:
  - Simple to use
  - Standardized representation
  - Fast lookup
- **Limitations**:
  - May not capture all clinical nuances
  - Limited specificity for complex cases

Post-coordinated Expressions

- **Definition**: Combination of multiple concept identifiers for complex meaning

- **Structure**:

  ```
  Primary Concept : Attribute = Value, Attribute = Value
  ```

- **Components**:

  - `Primary Concept`: Main clinical idea
  - `Colon ( : )` : Indicates refinement
  - `Attributes`: Qualifying characteristics (comma-separated)
  - `Values`: Specific values for each attribute

- **Example Structure**:

  ```
  Fracture : Finding site = Femur, Severity = Complete
  ```

When to Use Each Method

- **Pre-coordinated**: Standard, common clinical concepts
- **Post-coordinated**: Complex, specific, or unusual clinical situations

---

5. Description Logic Foundation

Logical Framework

- SNOMED CT is built on formal description logic principles
- Enables:
  - Automated reasoning
  - Consistency checking
  - Logical inference
  - Quality assurance

Benefits of Description Logic

- **Computational**: Machine processing and validation
- **Clinical**: Supports clinical decision support systems
- **Interoperability**: Consistent interpretation across systems
- **Quality**: Identifies inconsistencies and gaps

---

6. SNOMED CT Axes (Historical Context)

Historical Development

- SNOMED CT evolved from earlier SNOMED versions
- Previous versions used multi-axial structure (3.5 axes mentioned)
- Current version maintains conceptual organization while improving logical structure

---

7. Practical Applications

Clinical Documentation

- Enables precise coding of:
  - Diagnoses
  - Procedures
  - Findings
  - Symptoms

Healthcare Interoperability

- Facilitates data exchange between:
  - Electronic Health Records (EHRs)
  - Healthcare institutions
  - Different countries/regions

Analytics and Research

- Supports:
  - Population health studies

  - Clinical research
  - Quality measurement
  - Epidemiological analysis

---

8. Implementation Considerations

System Integration

- Requires mapping to local terminologies
- May need customization for specific clinical domains
- Integration with existing workflow systems

User Training

- Clinicians need education on proper concept selection
- Understanding of pre- vs. post-coordination
- Awareness of synonym options

Quality Assurance

- Regular updates and maintenance
- Consistency checking
- Gap identification and resolution

---

9. Future Directions

Ongoing Development

- Continuous expansion of sufficiently defined concepts
- Enhanced logical definitions
- Improved multilingual support

Technology Integration

- Natural language processing applications
- Artificial intelligence systems
- Clinical decision support enhancement

Key Takeaways

1. **Concept Stability**: Clinical concepts remain constant while descriptions provide flexible human interface
2. **Hierarchical Organization**: IS-A relationships create logical taxonomy
3. **Expression Flexibility**: Both simple (pre-coordinated) and complex (post-coordinated) expressions supported
4. **Logical Foundation**: Description logic enables automated reasoning and quality assurance

5. **Ongoing Evolution**: Continuous improvement toward more complete logical definitions

---

Study Tips

- Focus on understanding the relationship between concepts and descriptions
- Practice identifying when to use pre- vs. post-coordinated expressions
- Understand the importance of the FSN suffix for concept classification
- Remember that many concepts are still "primitive" and require ongoing development

Phenotype Algorithms Using Electronic Health Records and Natural Language Processing

Overview

Source: BMJ 2015;350:h1885 - Liao et al.
Key Concept: Developing algorithms to accurately identify patients with specific diseases/phenotypes using EHR data, incorporating both structured data and natural language processing (NLP) of clinical notes.

Background & Motivation

Why EHR-Based Phenotype Algorithms?

- **Primary Driver**: Increasing use of EMRs creates opportunities for clinical/translational research
- **Applications**: Pharmacovigilance, genetic association studies, pharmacogenetics
- **Advantage**: Can assemble cohorts faster (12-18 months) vs. years for prospective studies
- **Challenge**: Need accurate phenotype identification from EMR data

Data Types in EMRs

1. **Structured Data** (readily available/searchable):

   - ICD-9/10 codes, CPT codes
   - Electronic prescriptions
   - Laboratory values
   - Vital signs

2. **Unstructured Data** (requires NLP):

   - Clinical notes (progress notes, discharge summaries)
   - Radiology reports
   - Pathology reports
   - Family/social history

Natural Language Processing (NLP) Fundamentals

What NLP Does

- **Core Function**: Computational method to extract information from text using linguistic rules

- **Process**: Breaks sentences → identifies parts of speech → applies linguistic rules → extracts meaning
- **Key Task**: Identifies "concepts" in clinical text (e.g., "atrial fibrillation" and "auricular fibrillation" = same concept)

Advantages of NLP

1. **Data Availability**: Captures information not in structured data or where structured data accuracy is low
2. **Systematic Linking**: Links multiple terms to single concepts (e.g., "tobacco," "pack-year," "cigarettes" → smoking)
3. **Standardization**: Uses databases like SNOMED CT and RxNorm to standardize terminology

Key Terminology Resources

Resource	Purpose	Example
UMLS	Unified Medical Language System - links standardized biomedical terms	Maps concepts to unique identifiers
SNOMED CT	Systematized Nomenclature of Medicine - organizes health terminologies	Body structure, clinical findings
RxNorm	Normalized clinical drug names	Links simvastatin to Zocor, combination pills

NLP Software Systems (Open Source)

- **cTAKES**: Apache clinical Text Analysis and Knowledge Extraction System
- **HITEx**: Health Information Text Extraction system

Methodology: i2b2 Approach

Step 1: Define Research Question

- **Example**: Rheumatoid arthritis genetic risk factors study
- **Goal**: High positive predictive value (PPV >90%) for adequate power
- **Consideration**: Balance between accuracy and sample size

Step 2: Create Sensitive Data Mart

- **Problem**: Low disease prevalence (≤1%) in general population limits PPV
- **Solution**: Screen for patients with ANY data suggestive of phenotype
- **Method**: Clinical domain experts determine screening components
- **Example**: Multiple sclerosis screen includes ICD-9 codes for MS, encephalitis, demyelinating diseases

Step 3: Develop Algorithm Variables

1. **Customized Dictionary**: Clinical experts create comprehensive term list
2. **Structured Mapping**: Convert terms to ICD-9, CPT codes, prescriptions, labs

3. **Negative Predictors**: Include phenotypes with similar presentations
4. **NLP Mapping**: Map terms to UMLS concepts and unique identifiers
5. **Data Integration**: Combine structured + narrative data

Step 4: Create Training Set

- **Selection**: Random sampling from data mart
- **Size Determination**: Based on number of variables and phenotype prevalence
- **Gold Standard**: Clinical experts manually review and classify patients
- **Criteria**: Use validated classification criteria when available (e.g., ACR criteria for RA)

Step 5: Develop Classification Algorithm

- **Method**: Adaptive LASSO penalized logistic regression
- **Output**: Probability score (0-1.0) for each patient
- **Formula Example**:

  ```
  Logit(probability of PA) = intercept − 0.16(sex) + 0.73 log(1 + (NLP
  PA)) + 0.88 log(1 + (ICD-9 PA)) + 0.63(NLP treatment) + ...
  ```

- **Threshold Setting**: Adjustable based on research goals (specificity 90-97%)

Step 6: Validation

- **Validation Set**: All predicted cases + 50% random patients from data mart
- **Blinding**: Reviewers blinded to algorithm results
- **Criteria**: Same classification criteria as training set

Key Performance Metrics

Primary Metrics

- **Positive Predictive Value (PPV)**: Accuracy of algorithm
- **Sensitivity**: Proportion of true cases identified
- **Specificity**: Proportion of true non-cases identified

PPV Formula

```
PPV = (sensitivity × prevalence) / [sensitivity × prevalence + (1-
specificity) × (1-prevalence)]
```

Trade-offs

- Lower specificity threshold → more patients classified → potential for better power

- Higher specificity → fewer patients but higher accuracy

NLP Impact: Results from i2b2 Studies

Performance Comparison

Phenotype	Algorithm Type	Sensitivity	PPV	Cohort Size
Crohn's Disease	Structured only	64%	98%	-
	Structured + NLP	72%	98%	-
Ulcerative Colitis	Structured only	60%	97%	4,183
	Structured + NLP	73%	97%	5,522
Multiple Sclerosis	Structured only	68%	94%	-
	Structured + NLP	78%	95%	-
Rheumatoid Arthritis	Structured only	51%	88%	3,046
	Structured + NLP	63%	94%	3,585

Key Findings

1. **Universal Improvement**: NLP improved ALL algorithms
2. **Mechanism**: Added independent predictive variables, increased sensitivity
3. **Greatest Impact**: When structured data accuracy is low (e.g., RA ICD-9 PPV = 19%)
4. **Moderate Impact**: When structured data already accurate (e.g., UC ICD-9 PPV = 64%)

Implementation Considerations

Required Team Members

1. **Clinical Investigator**: Domain expertise, phenotype definition
2. **Biostatistician**: Algorithm development, performance evaluation
3. **EMR Informatician**: Data extraction, healthcare system specifics
4. **NLP Expert**: Text processing, concept mapping

Resource Requirements

- **Infrastructure**: Research copy of EMR, secure servers, terabytes of storage
- **Time**: Mapping clinical terms to NLP concepts is rate-limiting
- **Expertise**: Specialized team with multidisciplinary skills

Limitations

1. **Resource Intensive**: Time and expertise for variable identification/extraction
2. **Mapping Complexity**: Clinical terms to NLP concepts requires significant effort

3. **System Specific**: May require adaptation for different EMR systems
4. **Limited Testing**: Method tested on defined diseases, not extensively on outcomes like drug response

---

Applications and Future Directions

EMR Research Platform Benefits

1. **Speed**: 12-18 months vs. years for prospective cohorts
2. **Rare Diseases**: Particularly valuable for uncommon conditions
3. **Biorepository Integration**: Links clinical + genomic data
4. **Hypothesis Testing**: Traditional clinical/genetic association studies
5. **Hypothesis Generation**: Phenome-wide association studies (PheWAS)

Multi-institutional Applications

- **eMERGE Network**: Electronic Medical Records and Genomics
- **Portability**: Algorithms can be applied across institutions
- **Collaborative Studies**: Enable multicenter research

Practical FAQs

Institutional Requirements

- **EMR Capabilities**: Advanced EMR with billing codes, e-prescriptions, lab values, narrative notes
- **Technical Infrastructure**: Programmers, relational database expertise, secure research environment
- **Hardware**: Secure servers, terabytes of storage space

Team Assembly

- **Core Team**: Biostatistician, clinical researcher, EMR informatician, NLP expert
- **Communication**: Regular team meetings essential for multidisciplinary coordination
- **Common Ground**: Establish shared terminology (e.g., NLP "precision" = PPV, "recall" = sensitivity)

Key Takeaways

1. **NLP is Valuable**: Consistently improves algorithm performance across all phenotypes
2. **Team Approach**: Requires multidisciplinary collaboration with regular communication
3. **Context Matters**: NLP impact greatest when structured data accuracy is poor
4. **Flexibility**: Probability-based approach allows threshold adjustment for different research goals
5. **Validation Critical**: Always validate with expert review on independent dataset
6. **Resource Planning**: Significant infrastructure and expertise requirements
7. **Future Potential**: Promising for clinical/translational research, particularly rare diseases

Best Practices

- Start with comprehensive screening to create sensitive data mart

- Include both positive and negative predictors
- Use validated clinical criteria when available
- Plan for iterative refinement of variable mapping
- Consider research goals when setting specificity thresholds
- Always validate with blinded expert review

Computing for Medicine - Quiz 1 Study Notes

1. FAIR Data Principles

FAIR stands for: **Findable, Accessible, Interoperable, & Reusable**

- **Findable**: Data should be easy to find by both humans and computers
- **Accessible**: Data should be retrievable by standard protocols
- **Interoperable**: Data should be compatible with other datasets and systems
- **Reusable**: Data should have clear licensing and metadata for reuse

---

2. Interoperability Concepts

Definition

Interoperability: The ability of two or more systems or components to exchange information and use the information that has been exchanged.

Levels of Interoperability (Increasing Complexity)

1. **Syntactic** - Format/structure compatibility
2. **Semantic** - Meaning/content compatibility
3. **Process** - Workflow/business logic compatibility
4. **Human** - User interface/experience compatibility

Network Effect Formula

For **n hospitals** each with different standards:

- Number of inter-hospital mappings = **n(n-1)/2**
- Example: 12 hospitals = 12(11)/2 = **66 mappings**

---

3. Healthcare Data Standards

HL7 (Health Level 7)

- **Purpose**: Standard for exchanging healthcare information
- **Structure**:
  - Pipe-delimited messages (|)
  - First line always: **Message Header (MSH)**
  - Multiple segments per message allowed
  - Contains numbers, text, symbols within delimiters

- **Use**: Primarily for data exchange between medical equipment

ECG Data Standard

- **Standard Used**: **HL7** (for communicating ECG machine data)

---

4. FHIR (Fast Healthcare Interoperability Resources)

Core Components

- **Base-address**: Identifies a FHIR system service/server
- **Type**: Resource type
- **Id**: Unique identifier
- **URL**: Resource location

Mandatory FHIR Resource Elements

Required:

- Identifier
- Human Readable Summary
- Profile

NOT Mandatory:

- URL link

FHIR Resource Categories

a> Conformance Resources

- **Purpose**: Describe how a system does or should work
- **Examples**: ValueSet, Conformance, StructureDefinition

b> Administration Resources

- **Purpose**: Manage administrative side of healthcare
- **Examples**: Patient, Order, OrderResponse

c> Clinical Resources

- **Purpose**: Clinical summaries, record keeping, and planning
- **Examples**: Observation, Condition, CarePlan, AllergyIntolerance

d> Financial Resources

- **Purpose**: Support financial services in healthcare provision
- **Examples**: Claim, Coverage, ExplanationOfBenefit

Patient Resource

Purpose: Exchange demographic and administrative information of individuals receiving care or healthcare services

Profiles

Definition: Help constrain FHIR resources for specific use cases and specify restrictions

- Used when building FHIR apps for specific contexts (e.g., rural primary healthcare)

---

5. EHR Meta-Model Components

Functional Component

- **Includes**: Administrative discharge and billing features
- **Purpose**: Operational healthcare functions

Component Types

- **Organizational**: Structure and governance
- **Information System**: Technical infrastructure
- **Functional**: Operational features (discharge, billing)
- **Data Model**: Data structure and relationships

---

6. ABDM (Ayushman Bharat Digital Mission)

Mission

Create a digital platform for evolving the health ecosystem through wide-range of data, information and infrastructure services while ensuring security, confidentiality and privacy.

Milestones

- **M1, M2, M3**: Progressive implementation phases

Key Stakeholders

1. **Patients**
2. **Health Professionals**
3. **Health Facilities**
4. **Digital Solution Companies**

ABHA System

ABHA Address

- **Purpose**: Manage personal health records

- **Format**: `computingformedicine@abdm`
- **Requirement**: Must create account on HIE-CM (Health Information Exchange - Consent Manager)
- **Domain**: `@abdm` indicates which HIE-CM manages the address

ABHA Number

- **Format**: 14-digit unique number
- **Requirement**: Strong KYC verification required
- **Auto-generation**: Automatically available as `<14digitabhano>@abdm`

ABDM FHIR Stack Components

First Component for Patient Sign-up: **Health Information Providers (HIPs)**

Other Components:

- **Health Information Users (HIUs)**
- **Consent Manager (CMs)**
- **Gateway**

---

7. Ontologies and Knowledge Representation

Ontology Structure

Representation: **Directed Graph with Cycles**

- Allows for complex hierarchical and associative relationships
- Can have circular references and multiple inheritance paths

Ontology vs Knowledge Graph

Key Difference:

- **Ontology**: Generalized model explaining relationships between entities based on common properties (not individuals)
- **Knowledge Graph**: Data-specific, establishes links between actual data points
- **Relationship**: Knowledge Graph = Data + Ontology

Terminology Types

Index

Definition: List of relevant terms pulled directly from unstructured or semi-structured text

- Used for document indexing and search

Other Types

- **Ontology**: Formal representation of knowledge

- **Thesaurus**: Controlled vocabulary with synonyms
- **Terminology**: Systematic collection of terms

Biomedical Ontology Applications

Clinical Decision Support Benefits:

- Computational graphs for querying parent/child terms
- Synonyms, related terms, and preferred terms for biomedical entities
- Semantic standardization (not syntactic interoperability)

---

8. SNOMED CT Concepts

Key Characteristics

True Statements:

- Concepts are computational meanings that do not change
- Each concept has at least one "is a" relationship
- Relationship types stored as concepts with SNOMED CT IDs

False Statement:

- **Post-coordinated expressions**: Use MULTIPLE concept identifiers (not single) to represent complex clinical ideas
- **Pre-coordinated**: Use single concept identifier

SNOMED CT Structure

- **Concepts**: Stable computational meanings
- **Relationships**: "Is a" hierarchies and other semantic links
- **Descriptions**: Human-readable terms for concepts

Key Formulas & Numbers to Remember

1. **Interoperability Mappings**: n(n-1)/2 for n systems
2. **ABHA Number**: 14 digits with strong KYC
3. **FHIR Resource Categories**: 4 main types (Conformance, Administration, Clinical, Financial)
4. **Interoperability Levels**: 4 levels (Syntactic → Semantic → Process → Human)
5. **ABDM Milestones**: M1, M2, M3
6. **ABDM Stakeholders**: 4 main groups

Common Exam Traps & Clarifications

FAIR Principles

- **Not**: Fair, Flexible, Authenticated, Interactive, Responsive
- **Correct**: Findable, Accessible, Interoperable, Reusable

HL7 Messages

- **Can** have multiple segments per message
- **Cannot** only have one segment per message

FHIR Mandatory Elements

- URL link is **NOT** mandatory
- Profile **IS** mandatory

SNOMED CT Post-coordination

- Uses **multiple** concept identifiers, not single

Ontology Representation

- **Directed graph WITH cycles** (not without cycles)

Study Tips

1. **Memorize FAIR acronym** - Frequently tested
2. **Understand interoperability complexity order** - Syntactic < Semantic < Process < Human
3. **Know FHIR resource categories** and examples for each
4. **Remember ABDM key numbers** - 14-digit ABHA, @abdm domain
5. **Distinguish ontology vs knowledge graph** - Conceptual model vs data-specific
6. **Know HL7 message structure** - MSH first, pipe-delimited, multiple segments allowed
7. **Understand mapping formula** - n(n-1)/2 for interoperability between n systems

COMPUTING FOR MEDICINE - MOCK QUIZ 2

INSTRUCTIONS:

- Fill your credentials carefully
- Total marks: 30
- Time: 45 minutes
- Number of questions: 20

MULTIPLE CHOICE QUESTIONS (1 mark each = 17 marks)

1. In the context of EHR phenotype algorithms, what does PPV stand for and why is it crucial for genetic studies?

a> Patient Phenotype Validation - ensures accurate patient classification b> Positive Predictive Value - determines algorithm accuracy for statistical power c> Primary Prevention Variable - identifies risk factors for diseases d> Phenotype Processing Validation - confirms data quality standards

2. Which NLP task is most critical for extracting clinical concepts from unstructured EMR data?

a> Part-of-speech tagging of medical terms b> Named entity recognition and concept mapping c> Sentiment analysis of clinical notes d> Text summarization of discharge summaries

3. In FHIR terminology, what is the primary purpose of a "Profile"?

a> To authenticate users accessing FHIR resources b> To encrypt sensitive patient information c> To constrain and customize FHIR resources for specific use cases d> To monitor system performance and usage

4. The relationship between PPV and disease prevalence in EMR algorithms is:

a> Inversely proportional - lower prevalence leads to higher PPV b> Directly proportional - higher prevalence leads to higher PPV c> Independent - prevalence does not affect PPV d> Exponentially related - small prevalence changes cause large PPV changes

5. Which component of the ABDM framework is responsible for managing patient consent for health data sharing?

a> Health Information Providers (HIPs) b> Health Information Users (HIUs) c> Consent Managers (CMs) d> Gateway services

6. In SNOMED CT, post-coordinated expressions:

a> Use a single concept identifier to represent simple clinical ideas b> Use multiple concept identifiers to represent complex clinical ideas c> Are automatically generated by NLP systems d> Replace the need for pre-coordinated concepts

7. The adaptive LASSO penalized logistic regression method is used in EMR phenotype algorithms to:

a> Clean and preprocess unstructured text data b> Identify predictive variables and assign weights c> Validate algorithm performance on test datasets d> Convert clinical notes to structured data formats

8. Which database standardizes health terminologies and is essential for NLP concept mapping?

a> RxNorm only b> SNOMED CT only c> Unified Medical Language System (UMLS) d> ICD-10 classification system

9. In the i2b2 methodology, a "sensitive data mart" is created to:

a> Store patient data securely with encryption b> Include patients with any evidence of the target phenotype c> Remove patients with incomplete medical records d> Validate the accuracy of structured EMR data

10. The formula n(n-1)/2 in healthcare interoperability represents:

a> The number of patients needed for algorithm validation b> The computational complexity of NLP processing c> The number of mappings needed between n different systems d> The minimum sample size for phenotype studies

11. Which FHIR resource category would contain a "Coverage" resource for insurance information?

a> Conformance b> Administration c> Clinical d> Financial

12. Natural Language Processing improves EMR phenotype algorithms most significantly when:

a> The EMR contains mostly structured data b> Structured data accuracy for the phenotype is already high c> Structured data accuracy for the phenotype is low d> The patient population is very large

13. In HL7 message structure, segments are:

a> Always limited to one per message transmission b> Separated by comma delimiters c> Multiple segments allowed per message with pipe delimiters d> Encrypted for security purposes

14. The primary advantage of using probability thresholds in EMR algorithms over Boolean approaches is:

a> Faster computation time b> Better data security c> Adjustability based on research objectives d> Simplified implementation process

15. ABHA numbers in the ABDM framework:

a> Are 12-digit identifiers with basic verification b> Are 14-digit unique identifiers requiring strong KYC c> Can be shared among family members d> Are automatically generated without verification

16. Which type of EMR data typically requires manual medical record review without NLP?

a> ICD-9 diagnostic codes b> Electronic prescription data c> Laboratory test results d> Narrative clinical notes

17. In healthcare ontologies, the "is a" relationship represents:

a> Temporal sequences between medical events b> Causal relationships between diseases c> Hierarchical classification structures d> Geographic distribution of health conditions

---

SHORT ANSWER QUESTIONS

18. Explain the difference between pre-coordinated and post-coordinated expressions in SNOMED CT with one example each. (3 marks)

Answer: *[Space for student answer]*

---

19. Describe the validation methodology used in the i2b2 approach for EMR phenotype algorithms. Include the composition of the validation set and the importance of blinding. (5 marks)

Answer: *[Space for student answer]*

---

20. Compare and contrast the four levels of healthcare data interoperability (Syntactic, Semantic, Process, Human) with specific examples from healthcare systems. (5 marks)

Answer: *[Space for student answer]*

ANSWER KEY

Multiple Choice Answers:

1. b> Positive Predictive Value - determines algorithm accuracy for statistical power
2. b> Named entity recognition and concept mapping
3. c> To constrain and customize FHIR resources for specific use cases
4. b> Directly proportional - higher prevalence leads to higher PPV
5. c> Consent Managers (CMs)
6. b> Use multiple concept identifiers to represent complex clinical ideas
7. b> Identify predictive variables and assign weights
8. c> Unified Medical Language System (UMLS)
9. b> Include patients with any evidence of the target phenotype
10. c> The number of mappings needed between n different systems
11. d> Financial
12. c> Structured data accuracy for the phenotype is low
13. c> Multiple segments allowed per message with pipe delimiters
14. c> Adjustability based on research objectives
15. b> Are 14-digit unique identifiers requiring strong KYC
16. d> Narrative clinical notes
17. c> Hierarchical classification structures

Short Answer Answers:

18. SNOMED CT Expressions (3 marks)

- **Pre-coordinated**: Single concept identifier represents a complete clinical idea
  - Example: "Pneumonia" = single SNOMED CT code
- **Post-coordinated**: Multiple concept identifiers combined to represent complex clinical ideas
  - Example: "Pneumonia" + "caused by" + "Streptococcus pneumoniae" = multiple codes combined

19. i2b2 Validation Methodology (5 marks)

- **Validation Set Composition**: All patients classified with the phenotype + additional 50% random patients from data mart
- **Review Process**: Clinical domain experts manually review all patients using same criteria as training set
- **Blinding**: Reviewers are blinded to algorithm classification results to prevent bias
- **Performance Estimation**: Algorithm performance calculated based on validation set results
- **Importance**: Provides unbiased assessment of algorithm accuracy and generalizability

20. Healthcare Data Interoperability Levels (5 marks)

- **Syntactic**: Format and structure compatibility
  - Example: HL7 message format standards, XML vs JSON data formats
- **Semantic**: Meaning and content compatibility

  - Example: SNOMED CT for standardized medical terminology, ICD coding systems
- **Process**: Workflow and business logic compatibility
  - Example: Clinical decision support workflows, care coordination protocols
- **Human**: User interface and experience compatibility
  - Example: Consistent EHR interfaces, standardized clinical workflows for staff

QUIZ 1 2025

Here is a detailed solution to Quiz 1, drawing upon the provided course material.

Computing for Medicine - Quiz 1 Solution

Max. Marks - 100

Multiple Choice Questions (MCQs)

1. **In a scenario where two hospitals exchange allergy information for a patient using a FHIR standard syntax, but one system interprets "penicillin allergy" as a drug intolerance instead of an allergy, what is the most likely issue?**

   - **Correct Answer: b) Semantic interoperability failure**
   - **Explanation:** This scenario highlights a **semantic interoperability failure** because, while the systems successfully exchanged data (syntactic interoperability), they did not understand the *meaning* of the exchanged information in the same way. One system misinterpreted "penicillin allergy" as a general "drug intolerance," indicating a lack of shared understanding of clinical terminology and concepts.

2. **Your friend is a medical intern who thinks that an electronic health record (EHR) is just a scanned PDF of medical reports. What will you tell them to correct their thinking?**

   - **Correct Answer: b) EHR is a computer-processable collection of a patient's health data with a standardized model**
   - **Explanation:** An **EHR is defined as a securely stored, computer-processable collection of a patient's health information** that supports ongoing healthcare and includes diverse data like consultations, tests, and prescriptions. It is much more than just a scanned PDF; it uses standardized codes and data formats to enable storage, exchange, and analysis by computers.

3. **Your clinician friend shared a dataset using a shared Google Drive link named "data_final_use_this.csv". You find that I. No fear of persistent identifier is given. II. Metadata is missing. III. The link often breaks when ownership changes. Which FAIR principle is MOST violated here?**

   - **Correct Answer: a) Findable**
   - **Explanation:** The **Findable** principle of FAIR data requires that data has a **persistent unique identifier** and rich metadata. The issues described—no persistent identifier, missing metadata, and a broken link—all directly violate the "Findable" principle by making the data difficult or impossible to locate and identify reliably.

4. **You are doing a study that needs to identify patients with diabetes from data of two hospitals. One hospital uses an electronic health record system that uses ICD-10 codes, another uses SNOMED CT. What technique is essential before data integration?**

    - **Correct Answer: b) Ontology mapping**
    - **Explanation:** To integrate data from systems using different coding standards like ICD-10 and SNOMED CT, **ontology mapping** is essential. Ontologies and terminologies provide standardized sets of terms and define concepts and relationships formally. Mapping them allows for common vocabularies that all systems can understand, ensuring semantic interoperability and accurate data integration.

5. **Simpson's paradox is a phenomenon where the overall group effects are opposite to that of subgroups. This is caused due to:**

    - **Correct Answer: b) Confounding variables**
    - **Explanation: Simpson's Paradox** occurs when a trend appears in different groups of data but disappears or reverses when these groups are combined. This paradox is typically caused by a **confounding variable** that is not accounted for when aggregating the data. For example, in vaccine efficacy data, age was identified as a confounder that led to Simpson's Paradox.

6. **A retrospective study design differs from a prospective study design because it looks at data that were collected prior to initiation of the study. What is NOT true as a limitation of this design?**

    - **Correct Answer: c) It is more expensive as compared to prospective studies**
    - **Explanation:** A **retrospective study looks back at existing records**. Limitations include high risk of selection/information bias, limited control over data quality, and difficulty establishing temporal relationships. However, retrospective studies are generally **less expensive** than prospective studies because they utilize already collected data, avoiding the costs and time associated with new data collection.

7. **Which of the following is an example of confounding?**

    - **Correct Answer: a) Coffee drinking is associated with lung cancer because coffee drinkers are more likely to smoke.**
    - **Explanation: Confounding** occurs when a hidden variable affects both the cause (exposure) and the effect (outcome), making it appear as if there's a direct link. In this example, smoking is the confounder: it is associated with coffee drinking and is a direct cause of lung cancer, thereby creating a spurious association between coffee drinking and lung cancer.

8. **Which is NOT true of human physiological systems? They:**

    - **Correct Answer: b) are highly optimal**
    - **Explanation:** Biological systems, including human physiological systems, are characterized as adaptive, self-organizing, constantly changing, oscillatory, far from equilibrium, sensitive, and **not always optimal**. Physical efficiency in one aspect (like perfect networks in lungs) can even make a system more fragile, demonstrating that they are not always perfectly optimal for health.

9. **In DAG representing Smoking → Lung Cancer – Asbestos Exposure, the variable Asbestos Exposure is:**

  - **Correct Answer: c) A confounder**
  - **Explanation:** In the context of studying the relationship between Smoking (exposure) and Lung Cancer (outcome), if Asbestos Exposure is a variable that is both associated with Smoking (e.g., certain occupations lead to both smoking and asbestos exposure) and directly causes Lung Cancer, then **Asbestos Exposure is a confounder**. A confounder influences both the independent variable (smoking) and the dependent variable (lung cancer), potentially distorting the observed relationship between smoking and lung cancer.

10. **A radiology AI system predicts tumor presence with high accuracy on training scans. But in practice, its performance drops because new scans are stored in a different file format and resolution. This is an example of:**

    - **Correct Answer: a) Interoperability challenge**
    - **Explanation:** The issue described, where the AI system fails due to differences in file format and resolution of new scans, is a **technical interoperability challenge**. For effective use, systems must be able to exchange and *use* information, which includes compatibility in data formats and structures.

11. **Which of the following is an example of structured healthcare data?**

    - **Correct Answer: c) Blood pressure readings**
    - **Explanation: Structured data** is organized in clear fields and tables. Blood pressure readings are typically recorded as numeric values in specific fields within an EHR, making them structured data. X-ray images and doctor's narrative notes are examples of unstructured data.

12. **A policymaker sees a rise in obesity rates alongside increased smartphone use and concludes that Smartphones cause obesity. This is an error due to:**

    - **Correct Answer: d) Confounding**
    - **Explanation:** This conclusion represents an error due to **confounding**. It's likely that other, unmeasured factors (confounders) like lifestyle, diet, or socioeconomic status are influencing both increased smartphone use and obesity, rather than a direct causal link between smartphones and obesity.

13. **We think about interoperability to enable high-fidelity exchange of data, meaning, workflow, and action. Workflows are a part of:**

    - **Correct Answer: c) Process Interoperability**
    - **Explanation:** The four levels of interoperability include **Process Interoperability**, which specifically refers to the ability of data to be used effectively within workflows or business logic. The phrase "meaning" refers to semantic interoperability, "exchange of data" to syntactic/technical, and "action" or "improving care" to human/clinical interoperability.

14. **Which of the following best explains why EHRs are better than paper records?**

    - **Correct Answer: b) They allow structured storage, interoperability, and analytics**
    - **Explanation:** EHRs provide significant advantages over paper records by offering **structured storage**, which facilitates data exchange and processing for **interoperability** and advanced **analytics**. This enables improved patient care, research, and streamlined workflows.

15. **What is the biggest challenge in using unstructured healthcare data?**

    - **Correct Answer: b) It lacks a predefined organization**
    - **Explanation:** Unstructured healthcare data, such as doctor's notes and radiology reports, is in **free text** or narrative form and **lacks a predefined organization**. This inherent lack of structure makes it challenging for computers to directly understand and process for analysis without advanced techniques like Natural Language Processing (NLP). While it can be digitized (e.g., typed text), extracting meaningful information from it is the core difficulty due to its disorganization.

Short-Answer Questions

1. What is an ontology? How is it different from a simple data dictionary? Explain with an example and create a small ontology graph depicting your understanding of any four medical terms using is-a, has-a process.

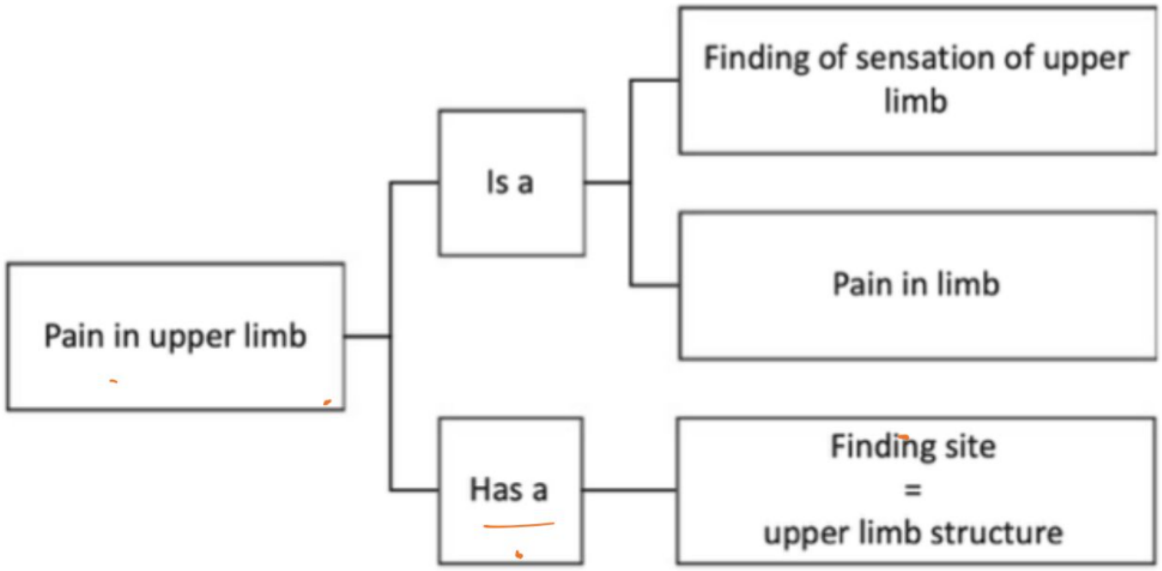
- **What is an Ontology?** An ontology provides both a **shared vocabulary** and the **rules or constraints for how those terms relate**. It formally defines concepts, attributes, and relationships, allowing computers to reason about data meaningfully. Ontologies use **description logic** to enable this reasoning, consistency checking, and logical inference. They are represented as a **directed graph with cycles**, allowing for complex hierarchical and associative relationships.

- **How is it different from a simple data dictionary?** A simple data dictionary is essentially a list of terms and their definitions, similar to a **controlled vocabulary** or **terminology**. It provides a standardized set of terms but typically lacks the formal logical structure, complex relationships, and reasoning capabilities of an ontology. An ontology, in contrast, is a **generalized model explaining relationships between entities based on common properties**, not just individual data points. A knowledge graph, for instance, is built using data and an ontology.

- **Example and small ontology graph:** Let's consider an example with four medical terms: "Disease," "Infectious Disease," "Pneumonia," and "Virus."

    - **Concepts:** Disease, Infectious Disease, Pneumonia, Virus
    - **Relationships:**
        - "Pneumonia" **is-a** "Infectious Disease"
        - "Infectious Disease" **is-a** "Disease"
        - "Pneumonia" **has-cause** "Virus" (or "caused-by")

```
        Disease
          ↑ (is-a)
 Infectious Disease
          ↑ (is-a)
        Pneumonia
       /         \
(has-cause)     (is-a type of)  -- (not explicitly in graph, but
implied by 'is-a')
       ↓
     Virus
```



  - **Explanation:** This graph shows a hierarchical "is-a" relationship, where "Pneumonia" is a specific type of "Infectious Disease," which in turn is a type of "Disease." It also shows a "has-cause" relationship, indicating that a "Virus" can cause "Pneumonia."

2. SNOMED CT concepts are more than just terms. They carry certain essential properties as discussed in class. Explain the essential properties of a SNOMED CT concept. What do you understand by Fully Specified Name (FSN)?

- **Essential Properties of a SNOMED CT Concept:**

  1. **Unique Machine-Readable ID:** Each concept in SNOMED CT has a unique identifier that is machine-readable, allowing for computational processing regardless of language.
  2. **Concept-Oriented (Constant Meaning):** SNOMED CT is fundamentally concept-oriented, meaning the clinical meaning of a concept remains constant and unchanging, independent of language or presentation.
  3. **Hierarchical Relationships ("Is-A"):** Every concept (except the root) has at least one supertype, forming an "is-a" hierarchy. This taxonomic structure allows for inheritance and classification, defining how concepts relate to broader categories.
  4. **Defined by Relationships:** Concepts are defined not only by their place in the hierarchy but also by **defining attributes** that specify additional characteristics, enabling precise automated inference for "sufficiently defined concepts".

- **Fully Specified Name (FSN):** The **Fully Specified Name (FSN)** is a type of description associated with a SNOMED CT concept. Its **purpose is to provide unique and unambiguous identification** of the concept. The FSN is not intended for end-users but rather for technical or administrative use. It follows a specific format, containing the concept name followed by a suffix in parentheses that indicates its primary hierarchy (e.g., "myocardial infarction (disorder)"). This suffix ensures precise identification of the concept across different systems and contexts.

3. Phrases such as "Possible fracture of arm" and "Patient is recovering well" are not valid SNOMED CT concepts. Why? Use a method to break them down using Post-coordinated valid concepts (need not be actual SNOMED-CT concepts).

- **Why they are not valid SNOMED CT concepts:** SNOMED CT concepts represent atomic clinical ideas, findings, procedures, or disorders. The phrases "Possible fracture of arm" and "Patient is recovering well" are more complex clinical expressions that convey **nuance, probability, or subjective assessment** rather than a single, atomic clinical fact that would typically be a pre-coordinated SNOMED CT concept.
    - "**Possible fracture of arm**": The term "possible" indicates uncertainty or suspicion, which is a qualifier of a finding rather than an intrinsic part of the "fracture" concept itself. SNOMED CT aims to represent concrete clinical facts.
    - "**Patient is recovering well**": This is a subjective assessment of a patient's status or progress, a narrative statement, rather than a specific medical condition, procedure, or finding defined as a singular concept. It encapsulates multiple ideas like "patient," "recovery," and "good progress."
- **Breaking them down using Post-coordinated valid concepts: Post-coordinated expressions** allow for the combination of multiple concept identifiers and attributes to represent complex, specific, or unusual clinical situations that are not covered by a single pre-coordinated concept.
    - **"Possible fracture of arm":**
        - **Primary Concept:** `Fracture (finding)`
        - **Attributes (with values):**
            - `Finding site (attribute)`:`Arm structure (body structure)`
            - `Certainty (attribute)`:`Possible (qualifier value)`
        - **Post-coordinated representation:** `Fracture (finding) : Finding site = Arm structure (body structure), Certainty = Possible (qualifier value)`
    - **"Patient is recovering well":**
        - **Primary Concept:** `Clinical finding (finding)`
        - **Attributes (with values):**
            - `Associated with (attribute)`:`Patient (person)`
            - `Progress (attribute)`:`Improvement (qualifier value)`
            - `Severity (attribute)`:`Good (qualifier value)`
        - **Post-coordinated representation:** `Clinical finding (finding) : Associated with = Patient (person), Progress = Improvement (qualifier value), Severity = Good (qualifier value)`

**4. You are building a predictive model for ICU patients. The data available include:

- Blood pressure readings
- Heart rate findings
- X-ray, CT and MRI images
- Text notes capturing patient history, progression and treatment**

a) Classify each of the above as structured, semi-structured, or unstructured. b) Which type of data is most directly usable for traditional statistical models, and why? c) What will be your approach to use unstructured data (e.g., notes) for predictive modeling?

- **a) Classify each of the above as structured, semi-structured, or unstructured:**

  - **Structured Data:**
    - **Blood pressure readings**
    - **Heart rate findings**
  - **Unstructured Data:**
    - **X-ray, CT and MRI images** (These are image files, rich in information but not organized in predefined fields for direct computational analysis like numeric data)
    - **Text notes capturing patient history, progression and treatment**
  - (No semi-structured data is explicitly listed in this scenario, though it could exist in other contexts.)

- **b) Which type of data is most directly usable for traditional statistical models, and why?** **Structured data** (Blood pressure readings, Heart rate findings) is most directly usable for traditional statistical models.

  - **Reason:** Structured data is organized in clear fields, tables, and typically consists of numerical or categorical values that are immediately computable. This format allows for straightforward application of statistical methods without extensive preprocessing or interpretation.

- **c) What will be your approach to use unstructured data (e.g., notes) for predictive modeling?** To use unstructured data like **text notes** for predictive modeling, the primary approach would be **Natural Language Processing (NLP)**.

  - **Process:** NLP involves computational methods to **extract information and meaning from text** using linguistic rules. This typically includes:
    1. **Named Entity Recognition (NER):** Identifying and classifying clinical concepts (e.g., diseases, symptoms, medications, procedures) within the free text.
    2. **Concept Mapping:** Linking the extracted terms to standardized terminologies and ontologies like **UMLS, SNOMED CT, and RxNorm** to ensure consistent interpretation and facilitate computational analysis. This standardizes various ways of expressing the same concept (e.g., "atrial fibrillation" and "auricular fibrillation" map to the same concept).
    3. **Feature Extraction:** Converting these standardized concepts and their relationships into structured features (e.g., presence/absence of a condition, frequency, severity) that can then be fed into machine learning models for prediction.
  - For **image data (X-ray, CT, MRI)**, the approach would involve **Computer Vision** or **Image Processing techniques** to analyze visual patterns and extract features relevant for the predictive model. These features, once extracted, can then be integrated with other structured data.