Computing for Medicine
Quiz - 1

Date - 28th August 2025 **Time - 60 minutes**

Max. Marks - 100 **Name:**

Multiple Choice Questions (MCQs) **4*15 = 60**

1. Two hospitals exchange allergy information for a patient using a FHIR standard syntax. One system interprets “penicillin allergy” as a drug intolerance instead of an allergy. What is the most likely issue?
   a) Human failure
   b) Semantic interoperability failure
   c) Process Interoperability failure
   d) Organizational failure

2. Your friend is a medical intern who thinks that an electronic health record (EHR) is just a scann deed PDF of medical reports. What will you tell them to correct their thinking?
   a) EHR is paperless storage
   b) EHR is a repository in computer-processable form with a standardized model
   c) EHR is a vendor’s private software for health data
   d) EHR is a software for administrative billing

3. Your clinician friend has shared a dataset using a shared Google Drive link named “data_final_use_this.xlsx”. You see that
   I. No DOI or persistent identifier is given.
   II. Metadata is missing.
   III. The link often breaks when ownership changes.

   Which FAIR principle is MOST violated here?
   a) Findable
   b) Accessible
   c) Interoperable
   d) Reusable

4. You are doing a study that needs to identify patients with diabetes from data of two hospitals. One hospital’s electronic health record system HR uses ICD-10 codes, another uses SNOMED CT.
   What technique is essential before data integration?
   a) Data encryption
   b) Ontology mapping
   c) De-identification
   d) Natural language processing

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

a. Data duplication

b. Confounding variables

c. Measurement error

d. Random noise

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?

a) It suffers from a high risk of selection bias and information bias due to reliance on existing records

b) Limited control over the quality and completeness of the collected data

c) It is more expensive as compared to prospective studies

d) It suffers from difficulty in establishing temporal relationships between exposure and outcome

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

a. Coffee drinking is associated with lung cancer because coffee drinkers are more likely to smoke.

b. A new vaccine reduces infection rates after randomization.

c. A measurement device consistently overestimates blood pressure.

d. Patients are lost to follow-up in a trial.

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

a) exhibit self-organizing behavior

b) are highly optimized

c) have non-linear properties

d) are adaptive

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

a) A parent node

b) A child node

c) A confounder

d) An exposure

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:

a) Interoperability challenge

b) Data leakage

c) Simpson's paradox

d) Random noise

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

a) X-ray images

b) Doctor's narrative notes

c) Blood pressure readings

d) MRI scans

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:
a) bad data
b) random sampling
c) overfitting
d) confounding

13. We think about interoperability to enable high-fidelity exchange of data, meaning, workflow, and action. Workflows are a part of:
a) Semantic Interoperability
b) Human Interoperability
c) Syntactic Interoperability
d) Process Interoperability

14. Which of the following best explains why EHRs are better than paper records?
a) They are cheaper to maintain
b) They allow structured storage, interoperability, and analytics
c) They require no security measures
d) They eliminate the need for doctors

15. What is the biggest challenge in using unstructured healthcare data?
a) It is always inaccurate
b) It lacks a predefined organization
c) It cannot be digitized
d) It is not useful for AI

Short- Answer Questions **10*4 = 40**

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.

   Ans. An ontology is a **hierarchical** graphical structure follows **dag** principle. formal, machine-processable specification of concepts within a domain and the relationships between them. In medicine, an ontology represents diseases, symptoms, procedures, drugs, and how they relate. For example, it can capture that "Diabetes mellitus is-a metabolic disorder" or that "Hypertension has-a measurement Blood Pressure." Ontologies are used in healthcare for semantic interoperability, decision support, and standardization of clinical knowledge across systems.
   How it differs from a Data Dictionary:
   A data dictionary is a flat list of terms with their definitions and metadata. Example: BP → Blood Pressure, unit = mmHg. An ontology adds much more: it shows **hierarchies** (is-a), **relationships, DAG** (part-of, has-a), attributes, and logical constraints. It allows reasoning. Example: From ontology, if "Hypertension is-a Cardiovascular disorder" and "All Cardiovascular disorders are Diseases," then a system can automatically infer that Hypertension is also a Disease.
   Example:
   Data dictionary: Glucose = blood sugar level in mg/dL.
   Ontology: Glucose (substance) is part of Metabolic process (process), measured in Laboratory Test (procedure)

Ontology Graph Example:

Patient
  has-a → Disease
    is-a → Diabetes
  has-a → Vital Sign
    is-a → Blood Pressure

Here, Diabetes is-a Disease, Blood Pressure is-a Vital Sign, and Patient has-a Disease / Vital Sign.
Definition of ontology (2 mark)
Difference from data dictionary (2 mark)
Example (1 marks)
Ontology graph with correct relationship (5 marks)

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

Ans.**Two essential properties of a SNOMED CT concept**:

a. **Unique Identifier (Concept ID):** Each concept has a unique, numeric, machine-readable identifier that never changes, ensuring consistency and avoiding ambiguity.
b. **Defined Meaning (Concept Definition):** Each concept has a precise, unambiguous meaning that is independent of the term used to describe it.

Other Essential Properties of SNOMED CT Concepts

c.Relationships (Concept Relationships): Each concept can be connected to other concepts using formal relationships (e.g., "Is a", "Part of", "Associated with"). This allows hierarchical organization (e.g., "Pneumonia" is a "Infectious disease"). A concept is **sufficiently defined** if its defining relationships are sufficient to distinguish it from all its supertype and sibling concepts.

d.Descriptions (Terms/Synonyms): Each concept has multiple human-readable descriptions.These include the Fully Specified Name (FSN), Preferred Term (PT), and Synonyms to support different clinical languages and practices.

e. Language Independence: The unique numeric Concept ID ensures that meaning does not depend on language. Descriptions can be translated into multiple languages without changing the underlying concept.

f. Versioning and Permanence: Concept IDs are permanent and never reused, ensuring long-term stability across different releases. (Each property: 2 marks × 2 = 4 marks)

Fully Specified Name (FSN): The Fully Specified Name (FSN) is the **unique, unambiguous description** of a **SNOMED CT** concept in **human-readable form**. The Fully Specified Name (FSN) is the most precise and complete description of a SNOMED CT concept in human-readable form. It ensures that each concept has a clear, unambiguous meaning that cannot be confused with similar terms. Every SNOMED CT concept is assigned one FSN per language or dialect, making it unique within that linguistic context. This uniqueness avoids ambiguity that may arise from common synonyms or abbreviations in clinical practice.

It provides the **exact meaning** of the concept, ensuring there is no confusion between similar terms. Each concept has **only one FSN** in each language/dialect. The FSN also includes a **semantic tag** in brackets at the end (e.g., *(disorder)*, *(procedure)*, *(finding)*). This tag specifies the category or domain of the concept, further clarifying its exact use in clinical documentation. For example: "**Myocardial infarction (disorder)"** is the FSN for the clinical concept commonly known as *heart attack.*

Thus, the FSN serves as the **gold standard description** of a SNOMED CT concept, ensuring consistency across electronic health records, languages, and healthcare systems.

Property of SNOMED - **4 marks**

Definition of FSN (clear, unambiguous description) → **2 marks**

Uniqueness (only one FSN per concept/language) → **2 mark**

Example (e.g., myocardial infarction, appendectomy) → **2 marks**

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

   Ans. **Why they are not valid concepts?**

   SNOMED CT concepts must be **clear, precise medical meanings**. SNOMED CT concepts are meant to capture ***clinical facts*** or **standardized clinical meanings**. Phrases like these introduce **uncertainty, temporality, or subjective interpretation**, which are not atomic medical facts. SNOMED CT does not have single pre-coordinated codes for vague/ambiguous phrases like "possible" or "recovering well." Instead, such phrases must be represented through **post-coordination** (combining multiple valid concepts with refinements). SNOMED CT concepts are meant to capture clinical facts or standardized clinical meanings. Instead, they should be expressed using **post-coordination**, where multiple valid concepts are combined. Phrases like *"Possible fracture of arm"* or *"Patient is recovering well"* are not valid because they mix The **disease/condition** (e.g., fracture) with **Extra information** such as "possible", "recovering", or "well". SNOMED CT does not allow such mixed phrases as a single concept. Instead, we need to break them into smaller valid concepts and then **combine them** (this is called **post-coordination**).

Example 1: ***"Possible fracture of arm"***

We break it into:

Fracture of arm (clinical finding)
Uncertainty / diagnostic certainty = possible (qualifier)

So in a post-coordinated way (pseudo-SNOMED):

Fracture of arm : 246090004|Associated finding| = 371930009|Fracture of arm|,

408729009|Finding context| = 415684004|Possible|

This means: There is a suspected/possible fracture of the arm, not a confirmed one.

Example 2: ***"Patient is recovering well"***

We break it into:

Recovery (process/clinical course)
Patient status = improving/recovering
Degree = well/satisfactory

Together → "**Patient in recovery phase with stable condition**".

So post-coordinated expression (pseudo-SNOMED):

Clinical course : 246090004|Associated finding| = 162467007|Recovering|,

363713009|Severity| = 255604002|Mild/Well|

This means: The clinical course of the patient is recovery, with a favorable progression.

Explanation of why phrases are invalid concepts → **4 marks**
Example 1: Breakdown of *"Possible fracture of arm"* into valid concepts → **3 marks**
Example 2: Breakdown of *"Patient is recovering well"* into valid concepts → **3 marks**

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

- blood pressure readings
- 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.

Blood pressure readings → Structured data
Because they are numeric values stored in tabular format (e.g., systolic, diastolic).
X-ray, CT, MRI images → semistructured data
Raw pixel values without fixed rows/columns like a database.
Text notes (history, progression, treatment) → Unstructured data
Free-text narrative without predefined format.

b) Which type of data is most directly usable for traditional statistical models, and why?

Structured data (blood pressure readings, lab values, vitals, etc.) are most directly usable. Because They are already in numeric/tabular form. Traditional models (like regression, logistic regression, Cox models) expect **rows = patients, columns = features**. No heavy preprocessing needed.

c) What will be your approach to use unstructured data (e.g., notes) for predictive modeling?

1. **Preprocessing text** - Clean notes (remove stopwords, handle abbreviations, standardize medical terms).
2. **Convert to numerical features-** Use NLP techniques:

    1. Bag-of-Words / TF-IDF (basic)
    2. Word embeddings (Word2Vec, GloVe)
    3. Contextual embeddings (BERT, ClinicalBERT, BioBERT)

3. **Feature integration -** Combine text-derived features with structured data (vitals, labs) to create a multimodal dataset.
4. **Modeling -** Use ML/DL models (XGBoost, random forest, or neural networks) that can handle high-dimensional features.
    (a) Classification → **3 marks**
    (b) Structured data explanation → **2 marks**
    (c) Approach for unstructured data → **5 marks** (stepwise explanation: preprocessing, feature extraction, integration, modeling)