جامعة المستقبل
AL MUSTAQBAL UNIVERSITY

كلية العلـوم
قسـم الأنظمة الطبية الذكية

المحاضرة الثالثة

Data Collection and Sources in Medical Simulation

المادة: Simulation and Modeling
المرحلة: الرابعة
اسم الاستاذ: م.م هادي صلاح هادي

Introduce the concept of data collection and its role in medical simulation

- Data collection is the process of gathering information from different sources to be used for analysis, modeling, and simulation.
- In Medical Simulation:
- Data represents real-world medical behavior (patients, diseases, treatments).
- It is the foundation for building realistic and reliable models.
- Without accurate data, simulations cannot produce valid results.
- Key Points:
- Data can come from sensors, medical devices, databases, or generated synthetically.
- Collected data is later preprocessed, analyzed, and used for modeling.

Figure: Medical data source

Why Data Collection is Important

1. Foundation of Simulation Models

- Simulation accuracy depends directly on the quality of collected data.
- poor data leads to unreliable models.

2. Realism in Medical Training and Research

- Good data makes simulations more realistic, helping doctors and students understand real patient behavior.

3. Supports Decision-Making

- Collected data helps predict disease outcomes, patient recovery, and treatment efficiency.

4. Enables AI and Machine Learning Integration

- Data is the fuel for machine learning algorithms that improve medical predictions and diagnoses.

5. Continuous Improvement

- More data = better model updates and calibration over time.

Scenario	Effect of Data
Accurate ECG data	Produces realistic heartbeat simulation
Noisy / missing data	Leads to incorrect diagnosis or model instability

Types of Medical Data

Type	Description	Example	Format / Source
Numerical Data	Continuous or discrete values that can be measured.	Blood pressure, temperature, glucose level	CSV, Excel, Sensors
Time-Series Data	Data recorded over time, showing trends or signals.	ECG, EEG, heart rate monitoring	.csv, .mat, .edf
Categorical Data	Data classified into groups or labels.	Gender, diagnosis type, disease stage	Database tables
Image Data	Visual information from scans or cameras.	MRI, CT, X-ray	.dcm, .jpg, .png
Textual Data	Written medical notes, prescriptions, or reports.	Doctor's notes, patient feedback	.txt, .pdf

Examples of Each Data Type

1. Time series

- **A time series** is data recorded over a period of time at regular intervals (daily, weekly, monthly…) to observe changes or trends.
- The image shows how brain activity changes over time.
- Left side: Colored brain areas show how active each region is.
- Right side: Line graphs show how the activity in these regions changes with time.

❖ **Why it's useful:**

1. Helps doctors and researchers see when the brain is active.
2. Shows which areas work together.
3. Helps understand brain function or find problems.

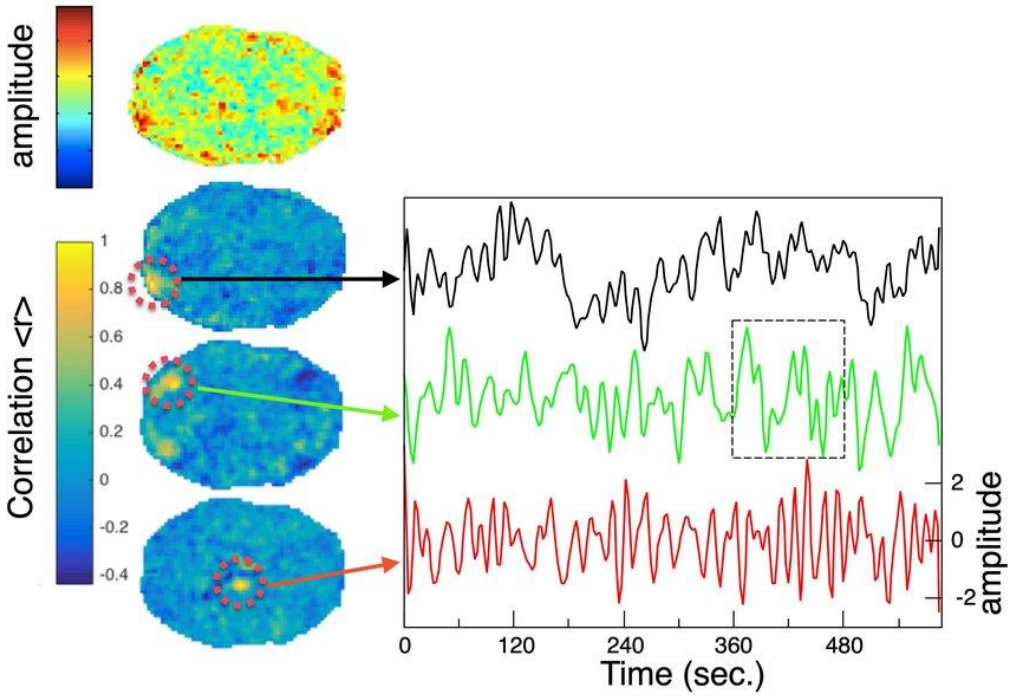


Figure: Brain signal → Time Series

2. Image Data:

Image Data: is visual information collected through cameras or scanners (like CT or MRI) and stored as digital images.

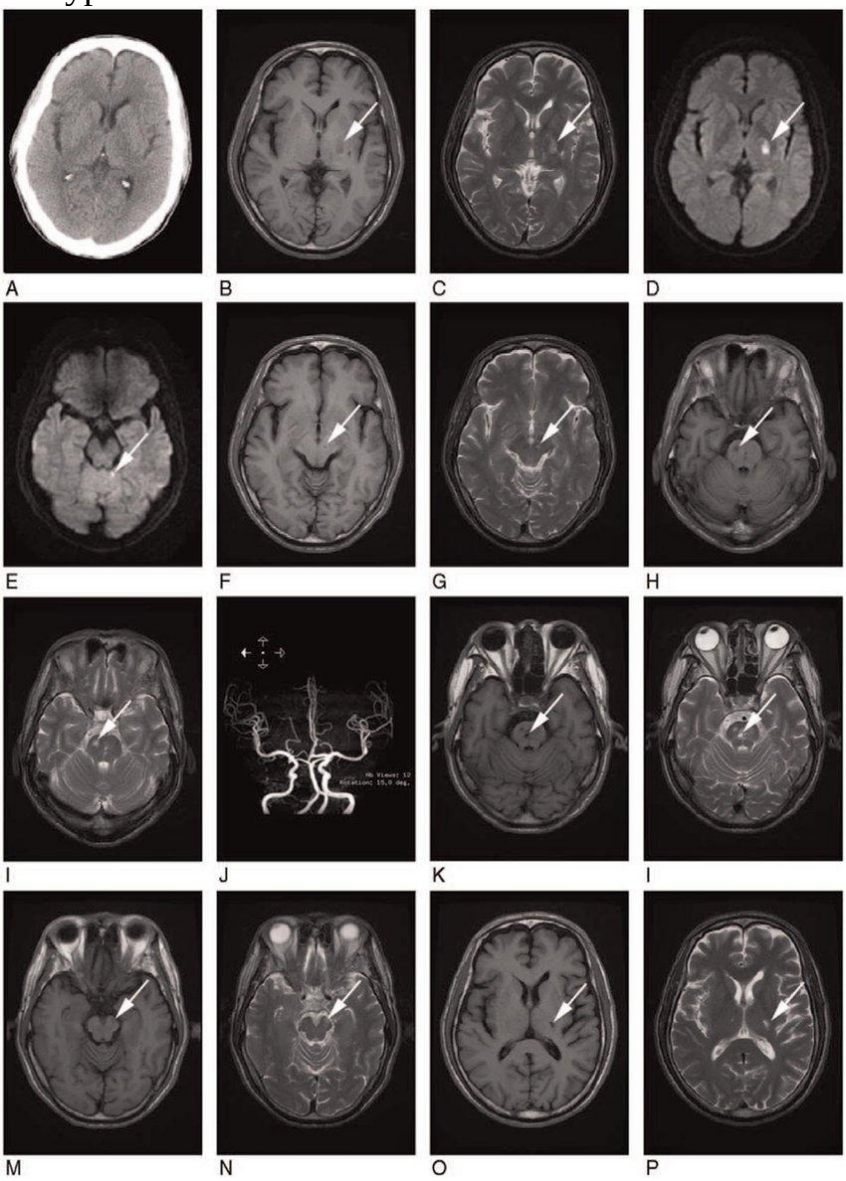
- Shows shapes, structures, or patterns.
- This image shows CT and MRI scans of the brain.
- It's an example of Image Data in medical analysis.

Why it's useful:

- Helps doctors see inside the brain.
- Detects abnormalities like stroke, bleeding, or tumors.
- Used for diagnosis and treatment planning.

Example:

- The arrows point to specific affected areas.
- Each scan type shows a different tissue detail



3. Numerical Data

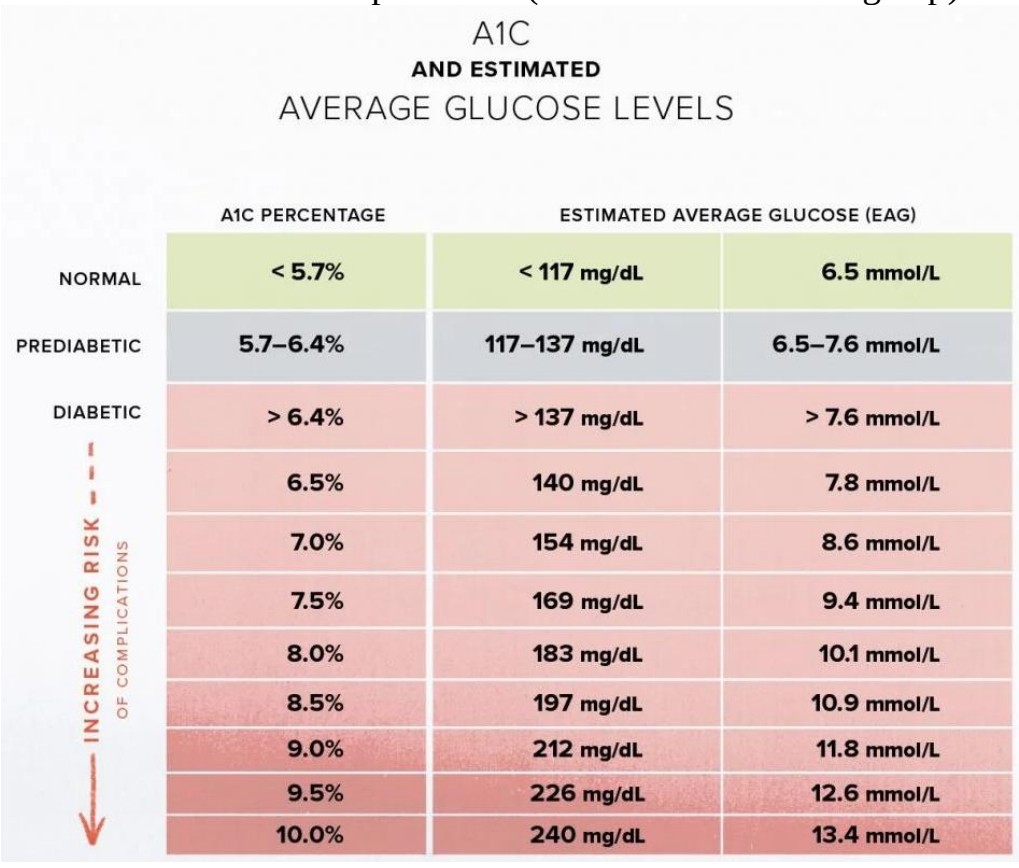
Numerical Data is data represented by numbers that can be measured or counted like blood sugar level, temperature, or heart rate.

About the image:

- It shows A1C levels (percentage) and estimated blood glucose (mg/dL and mmol/L).

These are numerical values used to:

- Check if blood sugar is normal, prediabetic, or diabetic.
- Measure risk level of complications (increases as numbers go up).



A1C
AND ESTIMATED
AVERAGE GLUCOSE LEVELS

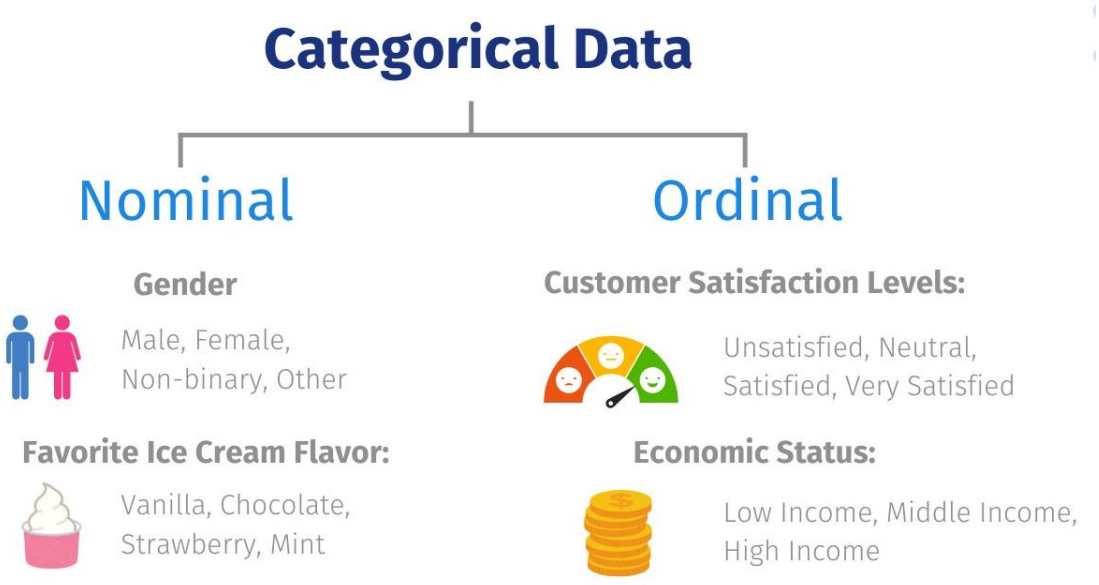
	A1C PERCENTAGE	ESTIMATED AVERAGE GLUCOSE (EAG)	
NORMAL	< 5.7%	< 117 mg/dL	6.5 mmol/L
PREDIABETIC	5.7–6.4%	117–137 mg/dL	6.5–7.6 mmol/L
DIABETIC	> 6.4%	> 137 mg/dL	> 7.6 mmol/L
INCREASING RISK OF COMPLICATIONS	6.5%	140 mg/dL	7.8 mmol/L
	7.0%	154 mg/dL	8.6 mmol/L
	7.5%	169 mg/dL	9.4 mmol/L
	8.0%	183 mg/dL	10.1 mmol/L
	8.5%	197 mg/dL	10.9 mmol/L
	9.0%	212 mg/dL	11.8 mmol/L
	9.5%	226 mg/dL	12.6 mmol/L
	10.0%	240 mg/dL	13.4 mmol/L

Figure: Glucose Chart → Numerical

4. Categorical Data

- **Categorical Data:** is data that describes groups or categories, not numbers.
- **Two types:**
- Nominal (no order):
- Examples: Gender (Male, Female), Ice cream flavor (Vanilla, Chocolate).
- Ordinal (has order):
- Examples: Satisfaction levels (Unsatisfied → Very Satisfied), Economic status (Low → High).

- From the image:
- Nominal data is about names or labels.
- Ordinal data is about ranked categories with a clear order.



Sources of Medical Data

Medical data can be collected from many different sources hospitals, devices, or open repositories and each source has its own strengths and challenges.

1. Clinical and Hospital Databases

- Real patient data collected during diagnosis and treatment.
- Examples: Electronic Health Records (EHRs), lab results, prescriptions.
- Format: .csv, .xlsx, or relational databases.
- Requires patient consent and data anonymization.

2. Medical Devices and Sensors

- Continuous monitoring devices collect real-time data.
- Examples: Heart rate monitors, glucose sensors, EEG/ECG devices, wearable watches.
- Format: time-series .csv, .mat, .edf.

3. Public / Open Datasets

- Freely available research datasets used for training and testing simulations.
- Popular sources:
- Kaggle: Medical datasets (MRI, ECG, patient data).
- PhysioNet: Physiological signal data (ECG, EEG, respiration).
- UCI Machine Learning Repository: Health data for ML projects.
- NIH / WHO databases: Global medical statistics.

4. Generated / Synthetic Data

- Artificially created data using statistical models or simulations.
- Used when real data is unavailable or restricted for privacy.

Data Representation for Simulation Concept:

- Before simulation, medical data must be stored and formatted correctly.
- Proper structure ensures that models can process data efficiently.

Example	Use Case	Format
Glucose, heart rate	Tabular / patient records	.csv
Signals, ECG	MATLAB simulation files	.mat
EEG/ECG data	Physiological signals	.edf
MRI, CT scans	Medical images	.dcm

```python
# Example of a small dataset creation
import pandas as pd
data = pd.DataFrame({
    "Patient_ID": [1, 2, 3],
    "HeartRate": [85, 90, 78],
    "Temperature": [36.7, 37.2, 36.5]
})
print(data)
```

Lab-2

```python
# Example showing how missing or wrong data affects analysis
import pandas as pd
data = pd.DataFrame({"Glucose":[90, 120, None, 300, 85]})
print("Before cleaning:\n", data.describe())
print("After filling missing values:\n",
data.fillna(data["Glucose"].mean()).describe())
```

Lab-3

```
# Example of mixed medical data
data = {
    "Patient_ID": [1, 2, 3],
    "Age": [45, 60, 52],           # Numerical
    "Gender": ["Male", "Female", "Male"],  # Categorical
    "HeartRate": [80, 76, 90],     # Time-series example
    "MRI_Scan": ["scan1.dcm", "scan2.dcm", "scan3.dcm"]  # Image
reference
}

df = pd.DataFrame(data)
print(df)
```