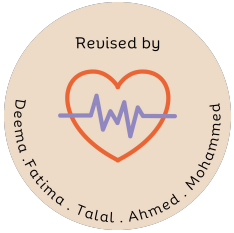




Informatics 442



References

Clinical data

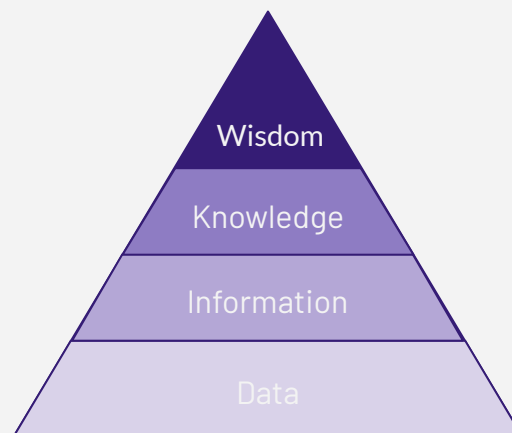
- Define data, information, and knowledge
- Types of clinical data
- Informatics vs. Information Technology and Computer Science
- Data to information
- Information to Knowledge
- Clinical Data Warehouse
- Use of Aggregated Clinical Data
- What Makes Informatics Difficult?
- Big data

Color index:

Main Text | **Female Slides** | **Male Slides** |
Extra | **Important** | **Dr's Notes (previous notes)** | **Dr's notes 442** | **Golden notes** |
Textbook



- **Data:** are symbols or observations reflecting differences in the world. (e.g. 5)
- **Information:** is data with meaning. (If we added the word "fingers" to number 5. "5 fingers", now it has a meaning)
- **Knowledge:** is information that is justifiably believed to be true. (elevation in FBG level, the likelihood of diabetes is increased. Smokers are likely to develop lung cancer more than non-smokers)
- **Wisdom:** is the critical use of knowledge to make intelligent decisions.



Bits can occur as various data types:

Each zero or one is a bit, each 8 bits equal a byte

- Integers (numbers ex: 14)
- Floating point numbers (decimals)
- Characters (A-Z)
- Character strings (ex: words)

File format:

Data are aggregated in many file formats, which makes sharing files possible.

- image files (JPG, GIF, PNG)
- text files
- sound files (WAV, MP3)
- video files (WMV, MP4)

Common or standardized file formats allow sharing of files between computers and between applications. For example, as long as your digital camera stores photos as JPG files, you can use any program that can read JPG files to view your photos.



Note that these data types do not define meaning. A computer does **not** "know" whether 3.14159 is a random number or the ratio of the circumference to the diameter of a circle (known as Pi or π).

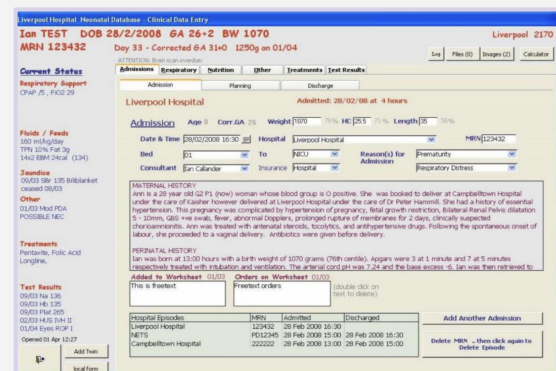
it is important to recognize that neither data types nor file formats define the meaning of the data.



Types of clinical data		
Narrative	Numerical measurements	Coded data
recording by clinician- maternity history	blood pressure, temperature , <u>glucose level, heart rate.</u>	selection from a controlled terminology system, <u>anything that has a drop down menu.</u>
Textual data	Recorded signals	Pictures
other results reported as text	EKG, EEG	radiographs, photographs, and other images

General categories of data entry:

- **Free-form** entry by historical methods:
 - Writing
 - Dictation
 - Typing
- **Structured (menu-driven) data** entry by mouse or pen.
- **Speech recognition** for either of above.



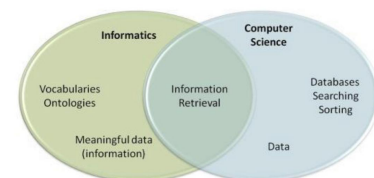
Informatics and Computer Science Important

What is the domain of computer science? DATA! (Meaningless)
 What is the domain of informatics? Information. (Meaningful)

Data are the domain of computer scientists and IT technologist
 , but information is the domain of **informatics** and informaticians
(The meaning of the data is of secondary importance to computer scientists).

-so we can say that computer science deals with meaningless data.

Information retrieval:
 involves both **computer science** (data) and **informatics** (information).



- Information retrieval is defined as finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need by retrieving documents from large collections (usually stored on computers).
- example of information retrieval task: find the relationship between heart attacks and aspirin > all records will be retrieved.

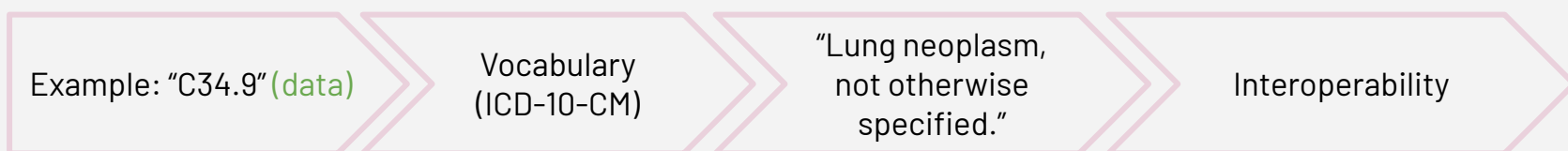
Informatics

- AI is concerned with the development of systems that can do something that previously required human intelligence. (chess, driving vehicles)
- **Example in medicine:** derma skin lesions categorization. This system was developed by Stanford university and it can categorize images of skin lesions to benign and malignant. This is done by training the system. Large data sets (images) are inserted with labels (benign, malignant,...). And then data sets (images) are added without labels.
- The system tries to learn to be able to identify lesions.
- Limitations:
 - 1- It requires large sets of labeled data to train the system.
 - 2- Cannot justify judgment (why did it classify this picture as benign?).

Data to information



Interoperability: Transmission of information between computer systems.



Explanation: C34.9 is code for lung cancer in computers. However, this Datum is of no meaning **if we didn't link it to a known vocabulary**. The computer still stores only data, not information

- **ICD:** International Classification of Disease, 10 is the version.
- **Interoperability:** Ability of two or more systems or components to exchange the information and use the information that has been exchanged. (ex: medical record at hospital A, sends information to a medical record at hospital B without losing this information).

Information to knowledge

Two types of data:

- **Unstructured (free text) such as discharge summaries or pathology reports.** (simple human language ex: history).
- **Structured data. may include billing codes, lab results (e.g., Sodium = 140 mg/dl).** (from drop down menus such as; billing codes, lab results, problem nests and medication nests).

- **Advantages and disadvantages :**
(ex: retrieval of all patients medical records that are in ICD-10-CM system with the same disease such as breast cancer).

- **Natural language processing (NLP):**
It's used to make computers understand unstructured data .
(Takes the natural human language as an input, process it (low level processing: assigns the part of speech to words "verb, noun" while high level processing such as Siri).

	Structured Data	Unstructured
Advantages	Much easier to manage.	able to express anything
Disadvantages	may not accurately reflect clinical reality	may be difficult to convey with a "one size fits all" vocabulary

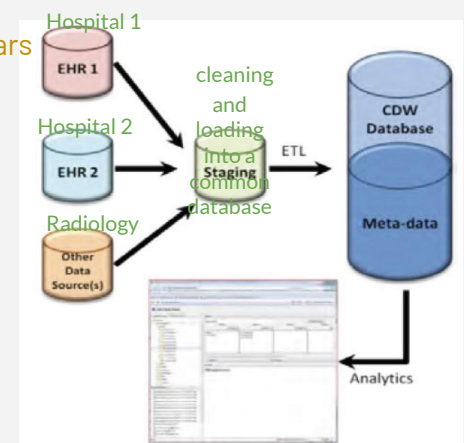
A modern way to convert medical information to knowledge is to use a clinical data warehouse (CDW).

Definition	a <u>database</u> system that collects, integrates and stores clinical data from a <u>variety of sources</u> including electronic health records (EHR), radiology and other information systems.
Staging	the process of cleaning the data from any incorrect or missing values.
Meta-Data	"data about data" example: ICD-10-CM
Difference between EHR&CDW	<ul style="list-style-type: none"> ● EHR: are systems implemented in hospitals that focus on individual data for one patient. <ul style="list-style-type: none"> -Address real time updates regarding patient data ● CDW: support many EHR, focus on group query. <ul style="list-style-type: none"> - Information update depends on how often it is updated (monthly, weekly,... unlike EHRs, they're designed to support real-time updating and retrieval of individual data

Uses

If you want to retrieve all women who are 40 years or older what will you use? CDWs

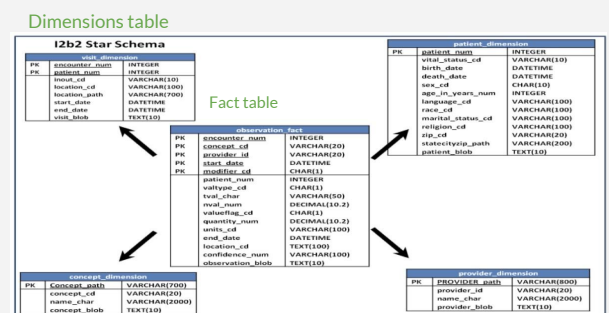
- **Monitor Quality** by allowing users to query for specific quality measures. e.g. retrieve all women who are 40 years or older who have not had a mammogram in the past year
- Identify trends.
- Comparative effective research(CER): (is the treatment A better than treatment B? We can answer this question by looking for the prognosis of people who were treated with both in the system)



i2b2¹ platform (an example of CDW)

- A harvard project to integrate biology and the bedside(i2b2) used by many other academic institutions in the US.
- Its an open source and modular and incorporate genomic and clinical information for research purposes.
- Database consists of **facts** (diagnoses, lab results, etc.) queried by users and dimensions that describe the fact.

- **Design: star scheme**
 - In the middle there is a fact table (quantitative, information).
 - On the side dimension tables (more information about these facts)
 - Dimensions tables provide qualitative information that answer (where, when and what)

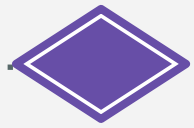


Use of Aggregated Clinical Data

- **Concept extraction**: the problem of identifying concepts within unstructured data, such as discharge summaries or pathology reports. غالباً يكون مربوط ب control vocabulary
 - Usually, these concepts are mapped to a controlled vocabulary.
- **Classification**: the problem of categorizing data into two or more categories.
 - Supervised machine learning. مثل سستم جامعة ستانفورد.

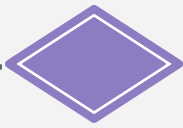
What Makes Biomedical Informatics Difficult?

Answer: "semantic gap" between data and information.



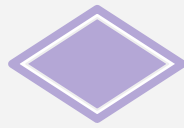
Incomplete information.

e.g.
unconscious patient in the ER. we don't know their past history.



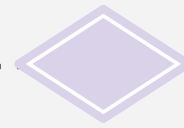
Uncertain information.

e.g.
عدد الأطفال بالسعودية
We can only estimate the answer using statistics.



Imprecise information.

e.g.
"pneumonia"
We need to know more
Viral? Bacterial?

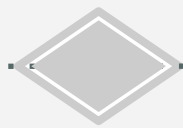


Vague information.

e.g.
Tall, big, elderly.



Vague information: information that includes quantifiers that permit boundary cases.



Inconsistent information
Ex: Birthdate:
8/29/66 and 9/17/66.

يعني احنا نعرف الفرق بين التاريخين بسيط وممكن نتغاضى عنه بس الكمبيوتر مايعرف



As illustrated in the example, all these imperfections may be found in healthcare. And humans can deal with these imperfections but computers cannot.

Big data

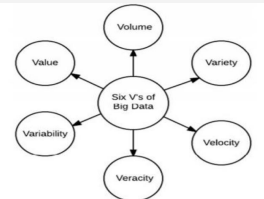
Definition: collecting large collections of data from various healthcare foundations followed by storing, managing, analyzing, visualizing, and delivering information for effective decision making.

The Six V's of big data:

It's important to know the six V's and their definitions



Click here to read more about big data.



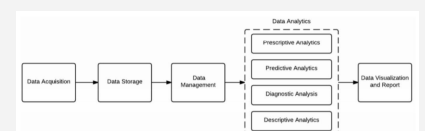
- **Volume:** large quantity of data produced by organization.
- **Variety:** a lot of sources in different formats, contribute in big data.
 - **Organized** (structured data: laboratory data, sensors data, data from databases).
 - **Semi-organized:** a combination of organized and unorganized (ex: website data and health records).
 - **Unorganized** "most common" (ex: social media data, medical records that doesn't use medical terminology).
- **Velocity:** massive frequency during the current details created, supplied and managed.
- **Veracity:** refers to accuracy of this data if it's accurate or valid (ex: in big data there is usually low veracity compared to other databases).
- **Variability:** data is fluctuating, changeable data, guidelines are changeable as well and might me updated so data might change during life.
- **Value:** methods of extracting valuable information from these data (big data analytics).



-Velocity includes both equivalent the rapidity of **data manufacture** and the rapidity of **data handling** to meet demand.
-Big data has low veracity, it can never be 100% accurate, and it is difficult to validate. Since most of the data comes from unknown sources.

Analysis: process of transforming raw data into information.

- **Stage 1 Data acquisition:** there are different sources (Electronic record, social media, smart phone, apps, websites... etc) to collect the data.
- **Stage 2 Storage:** store in cloud chambers.
- **Stage 3 Management:** organizing, cleaning and data mining to make sure that the data is valid, no missing data or values.
- **Stage 4 Analytics:**
 - Descriptive analysis: collects historical data, **what happened** in the health care?
 - Diagnostic Analysis: route cause of the problem, **why did it happen?**
 - Predictive Analysis: use both real time data and historical data, it has probabilities, **what will happen? what is the future trends.**
 - Prescriptive Analysis: number of different outcomes before the decision, what should we do? **And what is the best outcome?** And how we can make it happen?
- **Stage 5 Data visualization and reporting:** how will I represent these results? (ex: in graphs).



- Users must be able to “make sense” of clinical data; to make clinical data meaningful (data → information) and then learn from aggregated clinical data (information → knowledge)
 - Computer scientists focus on data, while informaticists focus on information
 - The transformation of information into knowledge is a primary goal of informaticists
- Clinical data warehouses are increasingly used to research clinical questions and generate knowledge from information
 - Sources of big data.



Key points

Extra

- Data are observations reflecting differences in the world (e.g., “C34.9”) while information is meaningful data or facts from which conclusions can be drawn and knowledge is information that is justifiably believed to be true
- Data are largely the domain of information technology (IT) professionals and computer scientists; information and knowledge are the domains of informatics and informaticians.
 - Vocabularies help convert data into information
 - The transformation of data into information and knowledge is a core concern of informaticians.
 - When the real world, the conceptual model and the computational model match, we get useful answers from the computer
 - Concepts relevant to health are relatively poorly defined compared to e.g. banking concepts
 - There is a large “semantic gap” between health data and health information

I encourage you to read this case study to understand the challenges at the information level and to understand the concept of interoperability.

Case Study: The Story of E-patient Dave

Extra

In January 2007, Dave deBronkart was diagnosed with a kidney cancer that had spread to both lungs, bone and muscles. His prognosis was grim. He was treated at Beth Israel Deaconess Medical Center in Boston with surgery and enrolled in a clinical trial of High Dosage Interleukin-2 (HDIL-2) therapy. That combination did the trick and by July 2007, it was clear that Dave had beaten the cancer. He is now a blogger and an advocate and activist for patient empowerment.

In March 2009, Dave decided to copy his medical record from the Beth Israel Deaconess EHR to Google Health, a personally-controlled health record or PHR. He was motivated by a desire to contribute to a collection of clinical data that could be used for research. Beth Israel Deaconess had worked with Google to create an interface (or conduit) between their medical record and Google Health. Thus, copying the data was automated. Dave clicked all the options to copy his complete record and pushed the big red button. The data flowed smoothly between computers and the copy process completed in only few moments.

What happened next vividly illustrated the difference between data and information. Multiple urgent warnings immediately appeared, including a warning concerning the prescription of one of his medications in the presence of low potassium levels (hypokalemia) (Figure 2.2). Dave was taking hydrochlorothiazide, a common blood pressure medication, but had not had a low potassium level since he had been hospitalized nearly two years earlier.

Worse, the new record contained a long list of deadly diseases (Figure 2.3). Everything that Dave had ever had was transmitted, but with no dates attached. When the dates were attached, they were wrong. Worse, Dave had never had some of the conditions listed in the new record. He was understandably distressed to learn that he had an aortic aneurysm, a potentially deadly expansion of the aorta, the largest artery in the human body.

Why did this happen? In part, it was because the system transmitted billing codes, rather than doctors’ diagnoses. Thus, if a doctor ordered a computed tomography (CT) scan, perhaps to track the size of a tumor, but did not provide a reason for the test, a clerk may have added a billing code to ensure proper billing (e.g., rule out aortic aneurysm). This billing code became permanently associated with the record. To put it another way, the data were transmitted from Beth Israel’s computer system to Google’s computer system quickly and accurately. However, the meaning of what was transmitted was mangled. In this case, the context (e.g., aortic aneurysm was a billing concept, not a diagnosis) was altered or lost. According to the definitions presented in chapter 1 (and reiterated later in this chapter), meaning is the defining characteristic of information as opposed to data.

After Dave described what happened in his online blog² (<http://epatientdave.com/>), the story was picked up by a number of newspapers including the front page of the Boston Globe.³ It also brought international attention to the problem of preserving the meaning of data. It became very clear that transmitting data from system to system is not enough to ensure a usable result. To be useful, systems must not mangle the meaning as they input, store, manipulate and transmit information. Unfortunately, as this story illustrates, even when standard codes are stored electronically, their meaning may not be clearly represented.



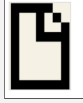
-Information is meaningful data or facts from which conclusions can be drawn (e.g., ICD-10-CM code C34.9 = "Malignant neoplasm of unspecified part of bronchus or lung").

-Knowledge is information that is justifiably believed to be true (e.g., "Smokers are more likely to develop lung cancer compared to non-smokers").

-Why do we aggregate data into file formats?

1-To specify the way the data are organized within the file.

2-Common or standardized file formats allow sharing of files between applications.



-When converting data to information, one must consider the **conceptual model** to convert it to the **computational model**.

Conceptual model: نموذج يهتم بالأشياء التي يريد أن يهتم فيها فقط، مثل نموذج تبخر الماء إلى سحاب إلى مطر، ويستتني عوامل أخرى مثل الرياح.

مثاله في الإنفورماتيكس: الاهتمام بالطول والوزن لبيانات أحد المرضى بغض النظر عن عمره. بعد ذلك، يمكن تحويل النموذج هذا إلى نموذج حاسوبي حسب الطلب

-The difference between **computer science** and **informatics** :

Computer science: computer scientists and IT professionals concentrate on **technology**, including computing systems composed of hardware and software as well as the algorithms implemented in such systems. For example, computer scientists develop algorithms to search or sort **data** (meaningless) more efficiently. Note that what is being sorted or searched is largely irrelevant. In other words, the meaning of the data is of **secondary** importance.

Informatics: Information and knowledge (meaningful), are addressed by informatics. To an informatician computers are tools for manipulating information.

-What is **information retrieval**?

Finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need by retrieving documents from large collections (usually stored on computers).

-Information retrieval is concerned with meaningful **information NOT data, and it involves computer science and informatics.**

-**AI** or "**machine learning**" have enabled computers to solve problems that have previously **resisted automation.**

-What is **deep learning**?

It's the use of multi-layer neural networks to learn patterns such as the features of objects in an image.

-Limitations of **deep learning**:

1-It requires large sets of labelled data to "train" the system. 2-The system cannot explain "why" it does something.

-**Structured** data vs **Unstructured** data

Structured:

-It may include billing codes, lab results (e.g., Sodium = 140 mg/dl), problem lists (**e.g., Problem #1 = ICD-10-CM C34.9 = "Lung Neoplasm, Not Otherwise Specified"**), medication lists.

-Structured data are much easier to manage and are computationally tractable.

Unstructured:

-Free text is simply human language such as English, called **natural language.**

-Key portions of clinical notes are still often dictated and are represented in records as free text.

-90% of data are unstructured.

-Has the advantage of being able to express anything that can be expressed using natural language.

-Difficult for computers to process as it requires NLP.

-**CDWs** vs **EHRs**

CDWs:

-Variety of analytics can be applied on data, and the results presented to the user via a user interface.

Examples of simple analytics include summary statistics such as counts, means, medians and standard deviations. More sophisticated analytics include associations (e.g., does A co-occur with B) and similarity determinations (e.g., is A similar to B).

-Designed to support queries about groups. -CDWs are used to identify trends.

EHRs:

-Designed to support real-time updating and retrieval of **individual** data.

-Databases that support EHRs are designed for efficient real-time **updating** and retrieval.

-**Semantic gap:** A wide semantic gap makes informatics difficult because concepts relevant to health are relatively poorly defined. e.g. "sick" can be high BP or low BP or an aortic aneurysm or any kind of illness.

Summary

Data

definitions

- **Data:** are symbols or observations reflecting differences in the world.
- **Information:** is data with meaning.
- **Knowledge:** is information that is justifiably believed to be true.
- **Wisdom:** is the critical use of knowledge to make intelligent decisions.

Clinical data

Types

- **Narrative**
- **Numerical measurements**
- **Coded data**
- **Textual data**
- **Recorded signals**
- **Pictures**

General categories of data entry

- **Free-form**
- **Structured (menu-driven) data**
- **Speech recognition**

Artificial intelligence (AI)

- **AI is concerned with the development of systems that can do something that previously required human intelligence.**

Data to information

- **ICD:** International Classification of Disease, 10 is the version.
- **Interoperability:** Ability of two or more systems or components to exchange the information and use the information that has been exchanged.

Clinical Data Warehouse

Definition

A modern **way to convert medical information to knowledge** is to use a clinical data warehouse (CDW).

Function

a database system that collects, integrates and stores clinical data from a variety of sources including electronic health records (**EHR**), radiology and other information systems.

Uses

- Monitor Quality by allowing users to query for specific quality measures
- Identify trends
- Comparative effective research: practice based research, answers very specific questions

Use of Aggregated Clinical Data

- **Concept extraction:** the problem of identifying concepts within unstructured data, such as discharge summaries or pathology reports.
 - Usually, these concepts are mapped to a controlled vocabulary.
- **Classification:** the problem of categorizing data into two or more categories.
 - Supervised machine learning.

What Makes Biomedical Informatics Difficult?

- Incomplete information.
- Uncertain information.
- Imprecise information.
- Vague information.
- Inconsistent information.

Big data

Definition

collecting large collections of data from various healthcare foundations followed by storing, managing, analyzing, visualizing, and delivering information for effective decision making.

MCQs

1-Finding material of unstructured nature that satisfies an information need by retrieving documents from large collections.

- A- Artificial intelligence
- B- Information retrieval
- C- Health informatics
- D-Machine learning

3-Designed to support queries about groups, and used to identify trends.

- A-Clinical Data Warehouse
- B-Electronic Health Record

5-Data that describe other data

- A- Meta-Data
- B-Big data

2-Key portions of clinical notes are represented as:

- A-Structured data
- B-Unstructured data

4-Big Data has low

- A-Velocity
- B-Veracity
- C-Variety
- D-Value

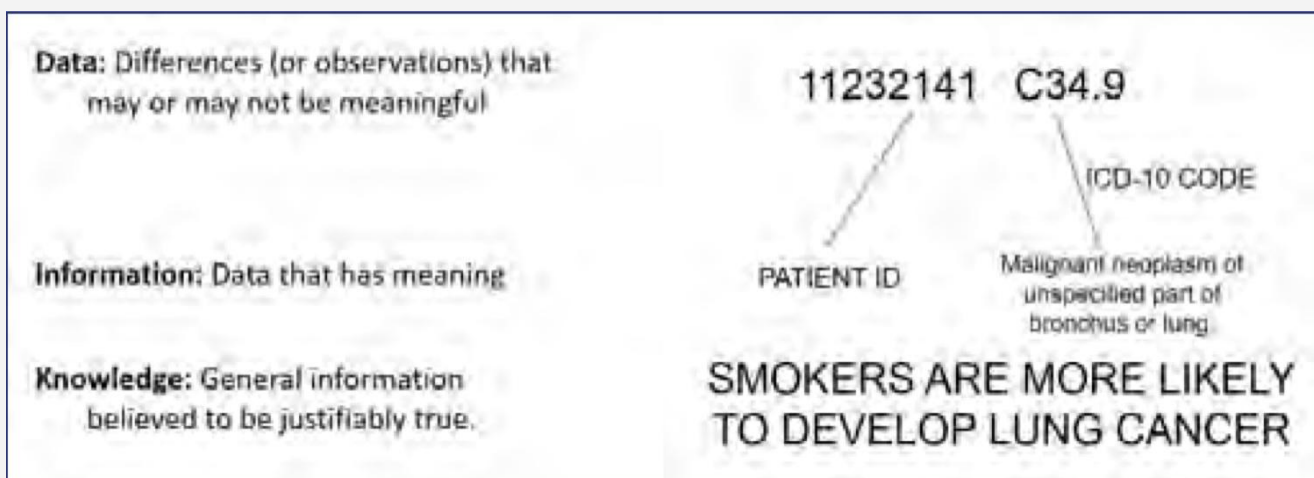
6-Which of the followings is picture data?

- A-ECG
- B-Blood test
- C-X-ray
- D-EMG

Answers key

1-B 2-B 3-A 4-B 5-A 6- C

Bonus!



Excuse the bad quality.

		n	%
Do you believe that COVID-19 exists	No	35	3.50%
	Yes	975	96.50%
What is the main reason for your willingness to receive the COVID-19 vaccine	To protect myself / my family from infection	869	86.00%
	To meet the requirements of work policies	113	11.20%
	To receive a permission to travel (vaccine passport)	28	2.80%
I have worries about getting COVID-19 infection after receiving the vaccine	Never	198	19.60%
	Occasional	238	23.60%
	Sometimes	374	37.00%
	Most of the time	112	11.10%
	Always	88	8.70%
After receiving COVID-19 vaccine, Adequate immunity against COVID-19 infection is obtained	Not getting enough immunity	76	7.50%
	Low probability	54	5.30%
	Average probability	354	35.00%
	High probability	368	36.40%
	Very high probability	158	15.60%
My risk perception towards COVID-19 has increased in comparison to before I received the COVID-19 vaccine	My perception hasn't changed	305	30.20%
	Low degree	60	5.90%
	Moderate degree	218	21.60%
	High degree	248	24.60%
	To a very high degree	179	17.70%
Risk perception scores	Mean	11	
	Minimum	3	
	Maximum	17	
	Standard Deviation	2	

Members Board



Informatics 442

Team leaders

Ghada Bin Slmah

Rand AlJarallah

Meshari Alshathri

Team members

deema aljuribah

Layan Aldoukhi

Fatima Halawi

Raneem Alwatban

Maram Beyari

Raid almadi

Amani Alotaibi

Nawaf Alrefaei

Sara Almajed

Fay aldossari

Hazem almalki

Abdulaziz Saad Alqahtani

Mohammed alzeer

