ULBS
Universitatea “Lucian Blaga” din Sibiu

Iceland
Liechtenstein
Norway grants

UNIVERSITETET I TROMSØ

ENSURE - Educating students for developing high quality research skills

This material was realised with the EEA Financial Mechanism 2014-2021 financial support. Its content (text, photos, videos) does not reflect the official opinion of the Programme Operator, the National Contact Point and the Financial Mechanism Office. Responsibility for the information and views expressed therein lies entirely with the author(s).

DATA COLLECTION, DATA CLEANSING, DATA VISUALIZATION AND EVALUATION

GABRIELA CÂNDEA

LUCIAN BLAGA UNIVERSITY OF SIBIU

DATA IN HEALTHCARE

- Generated across a variety of sources, data collection in healthcare can also encourage efficient communication between doctors and patients, and increases the overall quality of patient care

DATA COLLECTION

- **Data collection** is the ongoing systematic process of gathering, analysing and interpreting various types of information from various sources.
- **Data collection** in **healthcare** allows health systems to create holistic views of patients, personalize treatments, advance treatment methods, improve communication between doctors and patients, and enhance health outcomes

- Data is divided into two types:
  - **Quantitative** — in the form of numbers, e.g. percentages, comparison, etc.
  - **Qualitative** — in the form of words, e.g. description of quality, appearance, etc.

DATA COLLECTION

- The data collection instruments include
  - questionnaire surveys and patient self-reported data;
  - use of proxy/informant information;
  - hospital and ambulatory medical records;
  - and analysis of biologic materials.

DATA COLLECTION - QUESTIONNAIRE SURVEYS AND PATIENT SELF-REPORTED DATA

- The information collected in observational epidemiologic studies is collected in the form of patient/participant self-reports on standardized questionnaires which are either self or interviewer administered in person, by phone, or via mail or the internet

Advantages	Disadvantages
• Can collect personal and/or risk factor data not typically contained in hospital/ambulatory care records • Can elicit information in an analytically desirable and standardized manner • Can maintain high survey response rates through various financial or other incentives	• Validating individual survey responses can be difficult, burdensome, costly, and of questionable utility • If response rates are less than desirable, one may question the representativeness of the study sample and its generalizability • Responses might differ if questions are asked in-person vs. by phone vs. by mail/internet

DATA COLLECTION - PROXY/INFORMANT DATA

- The collection of information about study participants through the use of proxy respondents can be one of the more challenging tasks for an investigator.
- Informal caregivers are increasingly being recognized as ‘stakeholders’ in many research studies, particularly those that focus on patient reported outcomes such as quality of life.
- In cases of questionable mental status, or non-communicative state of a patient, informants can be very helpful and important in providing information to help establish a ‘baseline’ for a patient.

DATA COLLECTION - REVIEW OF AMBULATORY OR HOSPITAL MEDICAL RECORDS

- Information contained in hospital or ambulatory care records may be used either as the sole source of data, or complementary to other instruments used to elicit information.

Advantages	Disadvantages
• Readily available and contain much useful demographic and clinical information • Can be linked to other follow-up information sources • Can be used to characterize the medical history and clinical course of hospitalized and outpatient individuals • Can provide data on medication intensity and duration	• Often times data contained in medical records are non-standardized and inconsistently collected and recorded • Information is often incomplete and/or missing • Independent checks on validity and/or reliability are atypically performed • Information on etiologic or prognostic factors of importance is often either not obtained or asked about or recorded in a standardized manner

DATA COLLECTION - COLLECTION OF BIOLOGIC MATERIAL

- Contemporary clinical and translational research investigations involve the collection of biologic samples from study participants (such as hair, saliva, urine, and serum).
- Biologic samples are increasingly being used to profile participants metabolic, proteomic, or genomic status and, thereby, better understand their underlying pathophysiology or their response to a treatment or disease.

Advantages	Disadvantages
• May provide novel insights into underlying disease pathophysiologic processes • Can serve as an important endpoint of relevance • Can be linked to other sociodemographic, medical history, and clinical data to obtain insights into disease occurrence and prognosis	• Need to be collected under standardized conditions with considerable attention to detail • Ongoing quality control procedures needed • Need to consider impact of possible biologic circadian variation for purposes of timing and frequency of data collection efforts • May need collection of multiple measures at baseline to adequately profile subsequent changes

DATA CLEANSING

Referred to as **data** scrubbing, **data cleansing** is the process of detecting dirty **data** (**data** that is incorrect, out of date, redundant, incomplete or formatted incorrectly) and then removing and/or correcting the **data.**

Healthcare data tend to be unstructured for the simple reason that providers tend to format their data in whatever way is most convenient for them with no thought of the need to structure the data so that the cardiology department's data and the mental health department's data are consistent.

Don't blame the data! Clean it!

DATA CLEANSING

Dealing with data in healthcare:

- nowadays there is artificial intelligence which can diagnose cancer from an X-ray image with better accuracy than the best doctors.
- there are Internet of things (IoT) sensors within medical devices which can monitor the heartbeat of a heart attack survivor or the stomach of an individual with an ulcer to ensure their condition does not deteriorate.
- there is the new trend of healthy people using IoT devices to ensure that if they do get cancer or a narrowing of their arteries that they can go to the doctor's surgery before these illnesses develop into something much more severe

DATA CLEANSING - SOURCES

Healthcare data tends to come from lots of different places such as:

- from different specialities such as radiology, or
- from pharmacy, or
- from inpatients or outpatients, or
- from a GP's surgery or
- from a hospital.

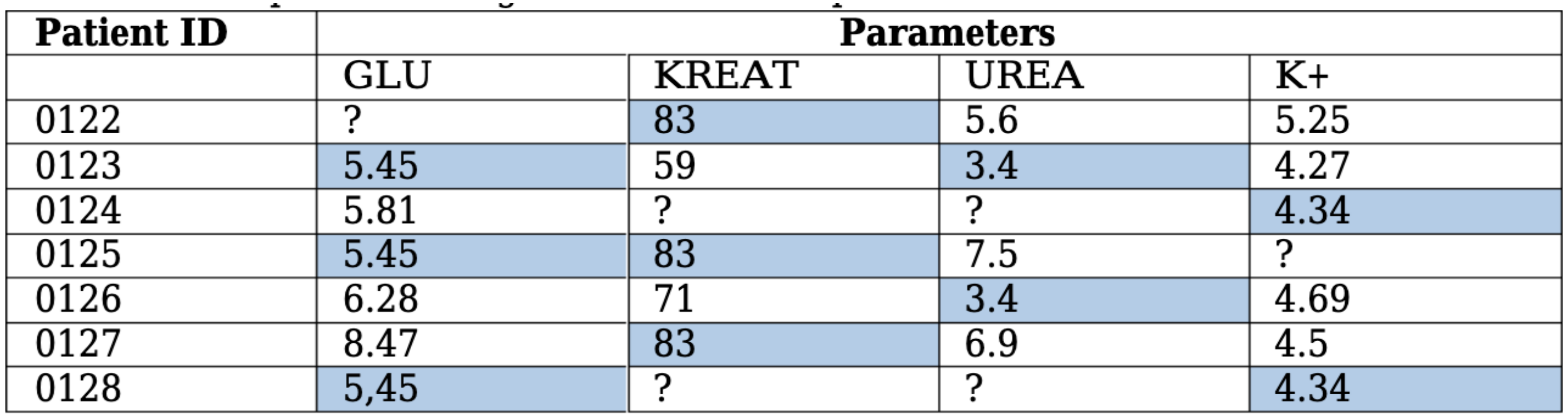
Data is typically stored in different ways and using different meta tags, also tend to come in different formats so that the radiology department's data will be full of X-rays, that is, images.

DATA CLEANSING - DELETING CASES / RECORDS WITH A MISSING ATTRIBUTE VALUES

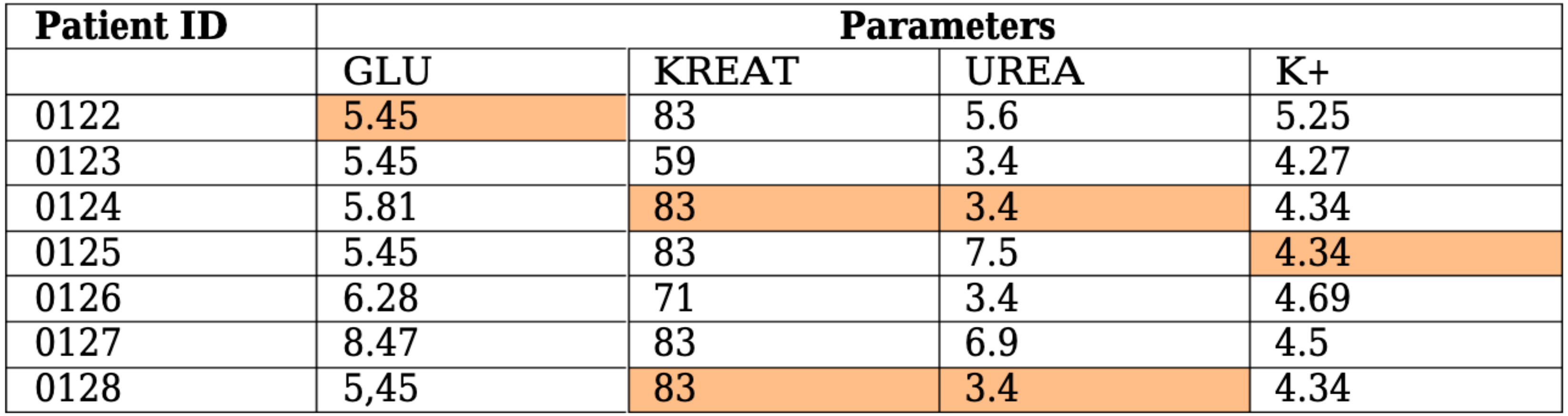
- All cases with missing values would be deleted from the data set.
- This approach is not appropriate for dealing with missing values in medical data.
- There are more than 40% of patients engaged in the research with missing attribute values.
- This approach will dramatically decrease the number of records which will cause the inability to perform data mining to a degree of accuracy as is required.

DATA CLEANSING - ADD MISSING ATTRIBUTE VALUES

Replacement of missing values by adding the most common value based on the values



Patient ID	Parameters			
	GLU	KREAT	UREA	K+
0122	?	83	5.6	5.25
0123	5.45	59	3.4	4.27
0124	5.81	?	?	4.34
0125	5.45	83	7.5	?
0126	6.28	71	3.4	4.69
0127	8.47	83	6.9	4.5
0128	5,45	?	?	4.34



Patient ID	Parameters			
	GLU	KREAT	UREA	K+
0122	5.45	83	5.6	5.25
0123	5.45	59	3.4	4.27
0124	5.81	83	3.4	4.34
0125	5.45	83	7.5	4.34
0126	6.28	71	3.4	4.69
0127	8.47	83	6.9	4.5
0128	5,45	83	3.4	4.34

DATA CLEANSING - DATA CLEANING PROBLEMS

Problem	Example

Detection of duplicate values	name1=“J.Kowalski”, born1=1978
	name2=“John Kowalski” , born2=1978

Detection of invalid values	name1= “J.Smith” , hight1=178
	name2= “J.Kowalski”, high2=”tall”

Detection of inconsistent values	born1= “1978”, age1=26, id1=781278xxxxx
	born2= “1960”, age2=30, id2=781260xxxxx

Detection of missing values	model1= ”Neo 02-Up”, serial_no1=5076096
	model2= ”LCP 201”, serial_no2=0
	model3= ”TRIOS 02-UP”,serial_no3=50464083

Detection of values out of scope	Id1=1, height1 =178
	Id2=2, height2 =-160

Separation of values embedded inside others	Name1 = ”Mr. John Smith” Name2 = ” J. Kowalski PhD”
	Forename1=”John”, surname1=”Smith”, title1=”Mr.”
	Forename2 =”J.”, surname2=”Kowalski”, title2=”PhD.”

DATA VISUALIZATION AND EVALUATION

- Data visualization is the process of analysing large amounts of data and communicating the results in visual context so that the audience can more easily digest and act upon the information.

DATA VISUALIZATION AND EVALUATION

Where can Data Visualization be used within health care?

- **Board presentations –** Provider and health presentations often provide an excellent opportunity to use Data Visualization techniques and data storytelling to present financial and other information like patient satisfaction results. Stories become more impactful when presented via Data Visualization, as opposed to solely using dull tabular columns.
- **Public health presentations –** Public health data, frequently vast and complicated, is easier to understand via Data Visualization.
- **Social determinants of health –** Less than 20% of mortality rates are affected by an individual's clinical care. Much of the remainder is based on sociological and economic risk factors (e.g., ethnicity, place of residence, education level, etc.). Healthcare systems and organizations are using Data Visualization to track and identify these social determinants of health in order to potentially positively influence them.

DATA VISUALIZATION AND EVALUATION - FREE DATA VISUALIZATION TOOLS

- https://informationisbeautiful.net/
- https://www.datawrapper.de/
- https://www.tableau.com/academic/students
- https://www.openoffice.org/download/