Stanford MEDICINE

Stanford | Center for Health Education

INTRO TO CLINICAL DATA STUDY GUIDE

MODULE 1 – ASKING AND ANSWERING QUESTIONS VIA CLINICAL DATA MINING

LEARNING OBJECTIVES

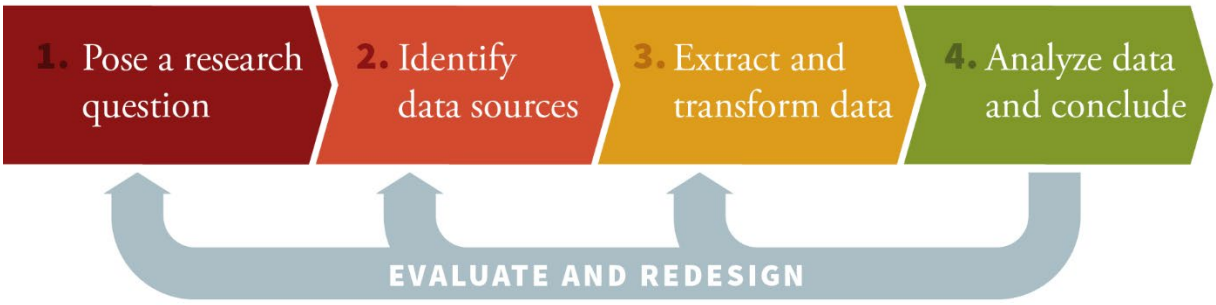
1. Explain the main steps in the data mining workflow
2. Describe the important categories of research questions
3. List properties that make a research question a useful one

THE DATA MINING WORKFLOW

Goal in this course is to explain how clinical data can be used to answer research questions to improve the health of patients and populations.

What we'll cover

1. How to choose research questions that are important
2. Structure of the healthcare system to understand how (and what) patient data are generated
3. Different kinds of data
4. Overview of the processing and analysis methods that get us answers to our questions
5. Problems and biases that can arise as well as ways to manage them



Data mining work will be referred in this course which has four steps:

1. Pose a research question.
2. Identify one or more data sources that can answer the question.

3. Extract and transform the data into a form needed for the analysis.
4. Conduct the analysis using those data

After completing the steps, results are evaluated and repeat the process if necessary.

Two representations of healthcare data that we will focus on this course are:

1. a patient timeline
2. a patient-feature matrix.

REAL LIFE EXAMPLE

MEET LAURA

A teenager with systemic lupus erythematosus (SLE), proteinuria, pancreatitis and positive for antiphospholipid antibodies

Laura

- A teenager with a chronic disease called Systemic Lupus Erythematosus (SLE).
- Has a flare up of the condition and develops proteinuria (protein in the urine), pancreatitis (inflammation of the pancreas), and has antiphospholipid antibodies in her blood.
- She is at risk for developing a blood clot.

Step 1 in the data mining workflow:

- Our clinical question:

  "Should a teenager with SLE who develops proteinuria and antiphospholipid antibodies receive an anticoagulant medication?"

(X) Review medical literature

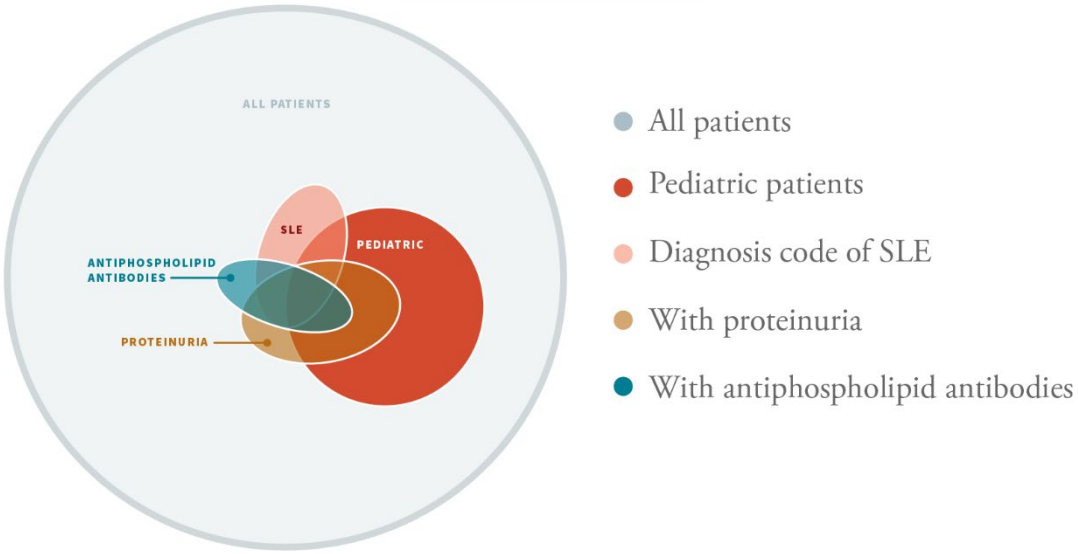
(X) Past experience

(X) Consult experts

- One approach is to examine what has happened to similar patients in the past, drawing on all the relevant data that appears in the electronic medical record, or EMR, of a large academic medical center.

Now we identified our data source, so we have completed Step 2 in the workflow.

EMR data are not necessarily organized in a way that makes searches straightforward. Medical expertise is needed to choose the diagnosis codes and medical terms that can identify patients that are in a similar situation.



Steps we would need to take.

1) Find all pediatric patients in the medical record system. Using a query based on patient age.
2) Diagnosis code of SLE.
3) Find patients with proteinuria by checking on the values listed in a urine test.
4) Find patients with antiphospholipid antibodies.
5) Laboratory test
    a) Recorded in numeric form

b) Result in textual document; Use treatment with aspirin as a proxy marker

c) Confirm the results of that search by checking laboratory results for those antibodies for those patients who have the results in a searchable form

6) Now in the midst of Step 3; we have defined our criteria that allows to identify the appropriate group of similar patients

7) Outcome of interest is clotting

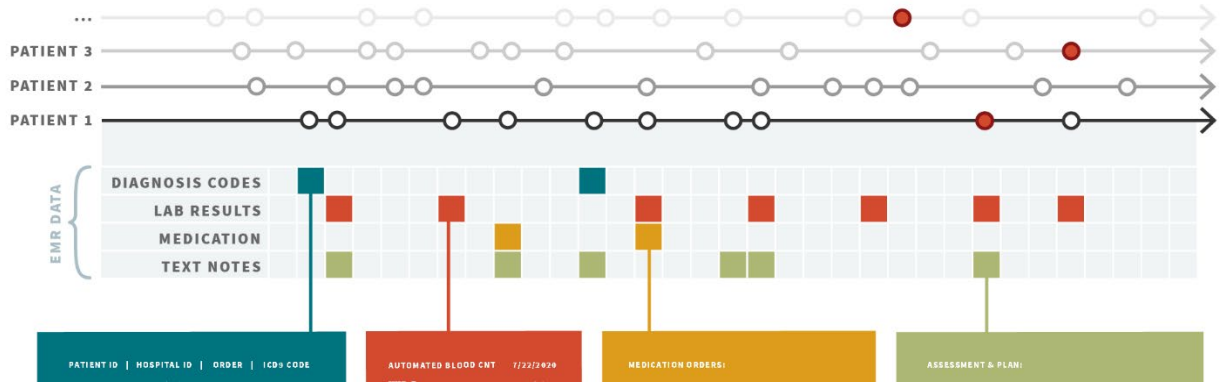
a) Search for "thrombus", "thrombosis", and "blood clot", again relying on clinical expertise to choose those terms.

Analysis is straightforward: compare the risk of clotting in teenagers with SLE, proteinuria, and antiphospholipid antibodies, to the baseline risk of clotting in teenagers with SLE.

- Need to get the patient data out of the EMR system in a form that allows analysis.
- Found that the relative risk of getting a blood clot when proteinuria and antiphospholipid antibodies are present is twice as high, when compared to the baseline, thus choosing to treat with an anticoagulant.

We have now completed Step 4 of the "data mining workflow".

Information in the form of **a patient timeline**.



- Start with data collected by the healthcare system, in this case, the electronic medical record, which includes diagnosis codes, lab results, medication orders, and text notes written by clinicians.

- Arrange these data on a timeline, one for each patient and then identify pediatric patients, and flag those with SLE, some of whom develop the comorbidities of concern (proteinuria, antiphospholipid antibodies).
- Use the timeline to count which patients experienced the outcome of clotting after they developed each clinical condition. Then compute the fraction of those with each condition who developed blood clots to arrive at the relative risk.

POSE A RESEARCH QUESTION	IDENTIFY DATA SOURCES	EXTRACT & TRANSFORM DATA	ANALYZE DATA & CONCLUDE
Is the risk of clotting in a teenager with SLE, proteinuria, and antiphospholipid antibodies high enough to warrant treatment with an anticoagulant?	Electronic Medical Record (EMR)	• Diagnosis codes • Medical expertise to craft searches for other terms • Proxy term in the search, followed by a second, confirmatory step.	Compared the risk of clotting in the subgroups to the risk of clotting for teenagers with SLE

Revisit the **data mining workflow** steps.

1. What was the clinical question? Is the risk of clotting in a teenager with SLE, proteinuria, and antiphospholipid antibodies high enough to warrant treatment with an anticoagulant?
2. What is the data source? The electronic medical record.
3. What extract/transform steps did we take? We defined how we will find teenagers with SLE, and how we will define subgroups based on clinical conditions. This involved the use of diagnosis codes in some cases and the use of medical expertise to craft searches for other terms. In one case we used a proxy term in the search, followed by a second, confirmatory step.
4. Finally, we compared the risk of clotting in the subgroups to the risk of clotting for teenagers with SLE in order to guide our decision to treat.

In this example we primarily used one data source, the EMR. Remember that there are no "perfect" ways of doing all the steps we reviewed in the example and it is best to think of the entire data

mining process as something that should be done with an expert human in the loop rather than by an automated algorithm that provides answers without knowing the larger context of the situation.

TYPES OF RESEARCH QUESTIONS

- A **descriptive question** asks for a summary of the data
- An **exploratory question** attempts to find what patterns might exist in the dataset available.
- An **inferential question** looks for patterns that go beyond just the particular dataset available. The goal is to find generalizable knowledge
- A **predictive question** looks for quantitative relationships between some features and the outcome of interest.
- A **causal question** looks for the effect of changes in one variable on a second variable.
- A **Deterministic question** is directly addressing the underlying mechanism

Clinical data are best suited for answering descriptive, exploratory, inferential, and predictive questions.

We ask these questions to accomplish two primary goals:

1. Risk stratification to decide if to treat
2. Data-driven selection of how to treat

What do you think about our analysis for Laura

The question we asked was about treating Laura who was experiencing a set of clinical conditions. However, the question we answered was about the proportion of patients with a set of clinical conditions who developed a blood clot. What we answered was a descriptive question relying on counts and proportions.

We then need to make an assumption that what happened in the past to those patients is likely to happen to Laura as well. The assumption, and the resulting conclusion, provides us with a 'risk-stratification'.

If we conclude that Laura is at high risk, what treatment to offer is clear, which is to use anticoagulation. In real life we would also need

to draw similar conclusions about the risks of adverse outcomes resulting from the treatment itself before making a final decision.

What makes answering a question useful:

- How many lives are affected? What is the disease burden?
- What is the chance that results will have a beneficial effect on the target community?
- What happens as a result of answering the question?
- Does knowing the answer help more than one constituent group among patients, healthcare professionals, and payers of care?