Stanford MEDICINE

Stanford | Center for Health Education

INTRO TO CLINICAL DATA STUDY GUIDE

MODULE 2 – DATA AVAILABLE FROM HEALTHCARE SYSTEMS

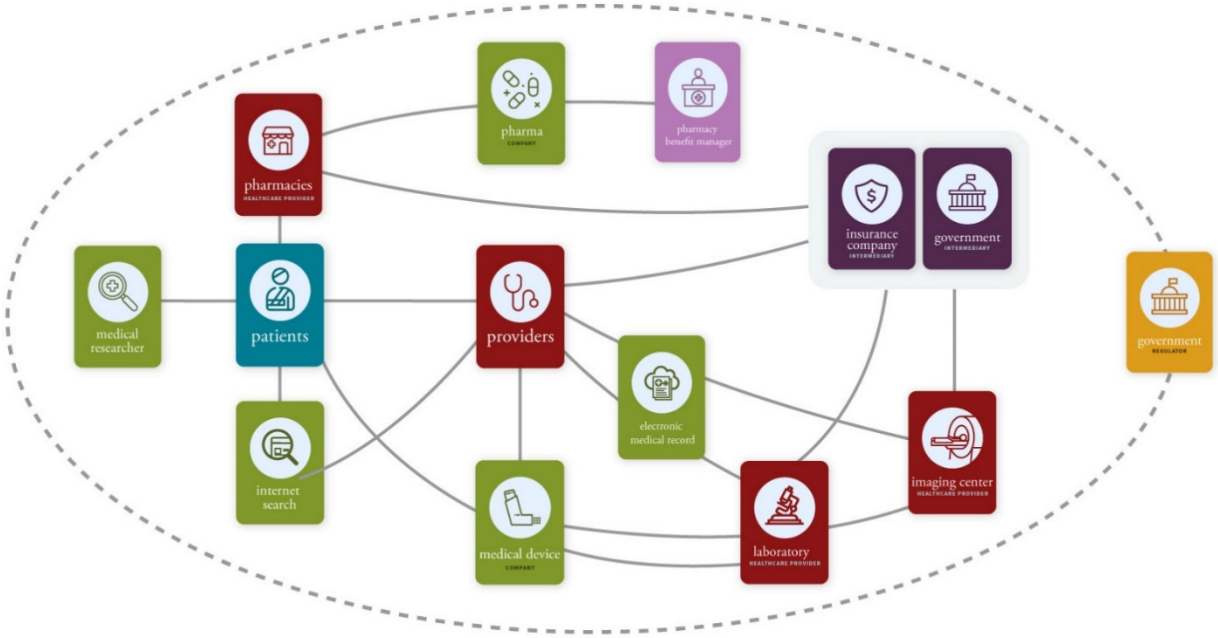
LEARNING OBJECTIVES

1. Describe the key actors in a healthcare system
2. Give examples of how different actors in the healthcare system can have different goals and interests.
3. List the different kinds of data the healthcare system produces
4. Describe the important healthcare data types
5. List pros and cons of using observational data
6. List examples of biases in observational data
7. Describe how to assess if a data source is useful

THE HEALTHCARE SYSTEM

Module goal is to show how clinical data can be used to ask and answer interesting and important research questions.

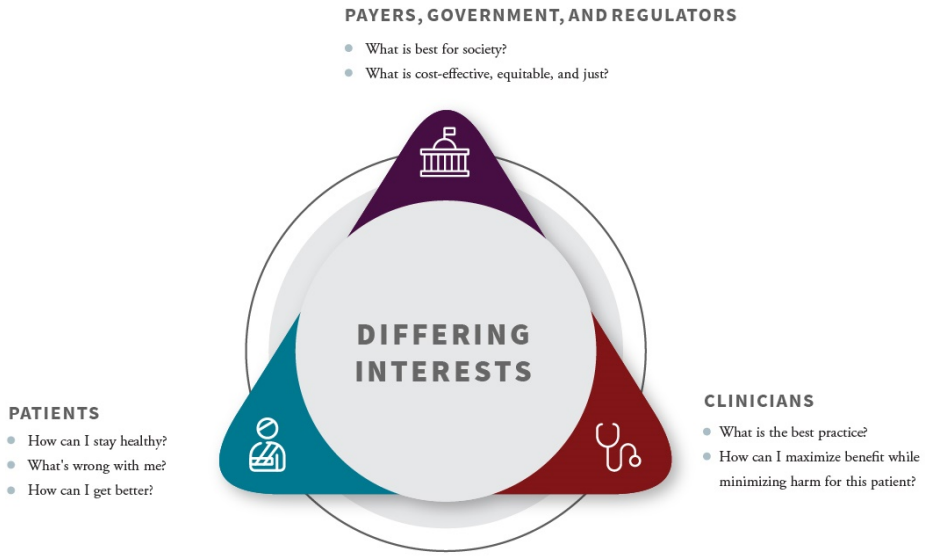
Below are the key entities in the healthcare system and how/what data each entity generates:



- **Patient:** If they need disease screening, or get sick, then they may seek treatment, or, more often search the internet to decide if they should seek treatment.
- **Healthcare providers:** Typically, the providers order laboratory tests or imaging procedures, make diagnoses, write prescriptions, and record their observations in the patient's electronic medical record (or EMR)
- **Pharmacies:** Provide drugs to patients
- **Pharmaceutical companies:** Design and manufacture drugs.
- **Drug distributors:** pharmacies procure drugs from this entity
- **Pharmacy benefits management companies:** manage payment for drugs
- **Medical device companies:** Design and manufacture medical equipment
- **State and federal government agencies:** Monitor and regulate the healthcare system
- **Governments:** Collect healthcare data to understand patterns of disease, which groups are underserved, and how healthcare is paid for
- **Medical researchers:** Investigate the patients, their disease conditions, and medical and surgical treatments. Publish their research results to increase scientific understanding of health and disease, and to guide the development of new drugs, devices, and other treatments. Also address policy questions on how we should best organize the provision of medical care.

The healthcare system is extremely complex, with many different types of actors, and many actors of each type. Actors can have different interests, which are not always in alignment.

The primary actors in the healthcare systems:



Different interests lead to tension in the delivery of healthcare, in the generation of data, and their use. It is important to keep these different interests in mind when defining a research question and choosing a data source to answer the question.

Think about:

- Who could benefit from answering your question
- How their interests could introduce some biases into the data sources used
- Ways to address more than one audience or interest group at the same time

HEALTHCARE DATA TYPES

Types of healthcare data:

- **Structured data:** Consistent organization; a table with rows and columns
- **Unstructured data:**
    - **Clinical text:** Quite different from ordinary written language, a haiku of acronyms
    - **Images:** Large two-dimensional arrays of intensity values, measuring the degree to which some kind of physical energy is transmitted or absorbed by tissue. Sometimes many two-dimensional images are collected together to form representations of volumes.
    - **Signals:** Measurements coming from a sensor, usually at regularly-spaced time intervals

Healthcare data vary along several dimensions: occur over different time scales, generated at different points in the patient's care journey, different possible values, different patterns of missing values

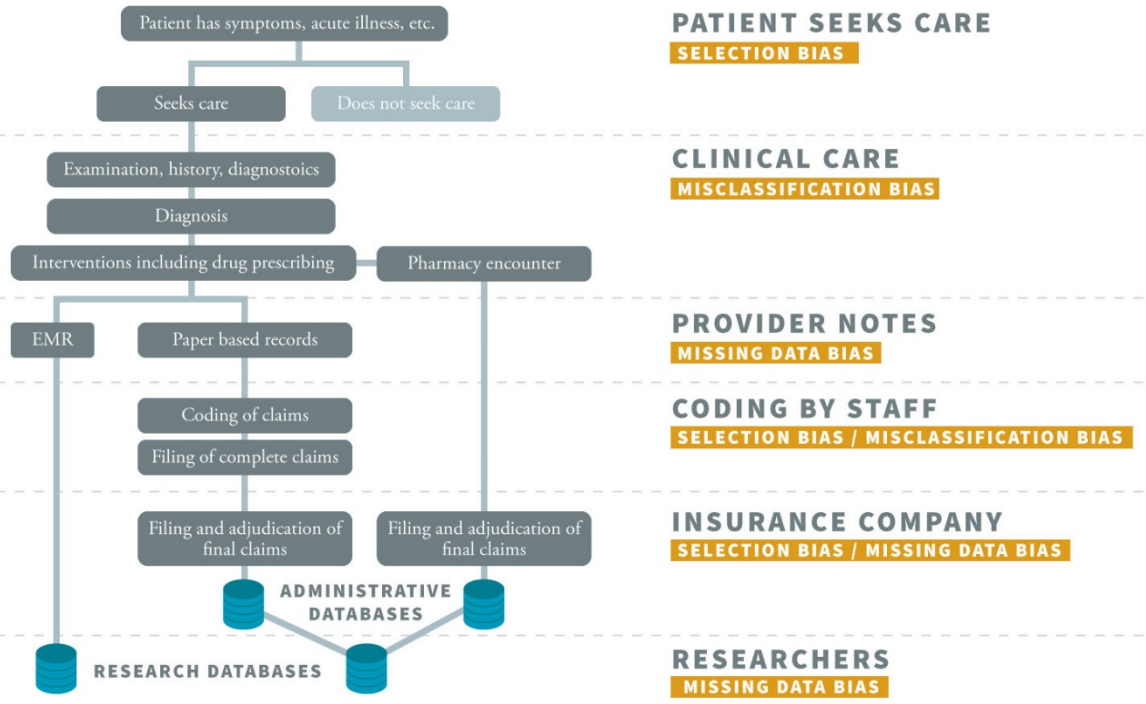
There is an **inherent time ordering** to all the different data types that are generated for a particular patient at various places in the healthcare ecosystem. Knowing this **timeline view of the data** is essential for effectively using the data.

- **Observational data:** data collected for other purposes and are collected as a byproduct of care delivery. Our use is referred to as *secondary use*.

Strengths of Observational Data	Weaknesses of Observational Data
• Very large datasets: Ability to study rare events	• Datasets are static: Difficult to obtain additional features/detail • Patient records not always linked

• Datasets are created during the routine operation of the healthcare system: Study real-world effectiveness and utilization • Available at relatively low cost without long delays: Accessible and efficient • Not subject to rapid changes in format, and the format may even be standardized	• Data creation and collection can be imperfect and biased, and may require significant cleanup • A record exists only if something (bad) happens

SOURCES OF BIASES AND ERRORS



Ways in which the data produced by each entity in the healthcare system might be inaccurate or biased:

- **Patient:** Decides to seek, or not seek, care. Many patient factors that might affect their decision, thus making this a selection bias
- **Clinical Care:** Not recording the health status of everyone, because leaving out those in their normal state of health. Not recording the health status of those who are sick but who treat themselves at home with over-the-counter medications, or those who are treated outside of

them health system for which we have records. In general, records are neither complete, nor a sample chosen at random

- **Healthcare Provider:** Might respond to financial incentives by changing how they decide whether to treat and which treatment to offer. Those incentives may also affect how they record what they did. Their records may be written at the time of service or with some delay afterwards. As a result, these records may be inaccurate or incomplete.
- **Coding:** Assign diagnosis and procedure codes to the medical documentation for the purpose of generating a bill for services rendered. A medical bill often contains only enough information to construct the bill, not a complete record of the treatment or what happened as a result of the treatment. There may be biases from systematic errors in coding. Coders may make mistakes as well as omit codes that are unlikely to be reimbursed.

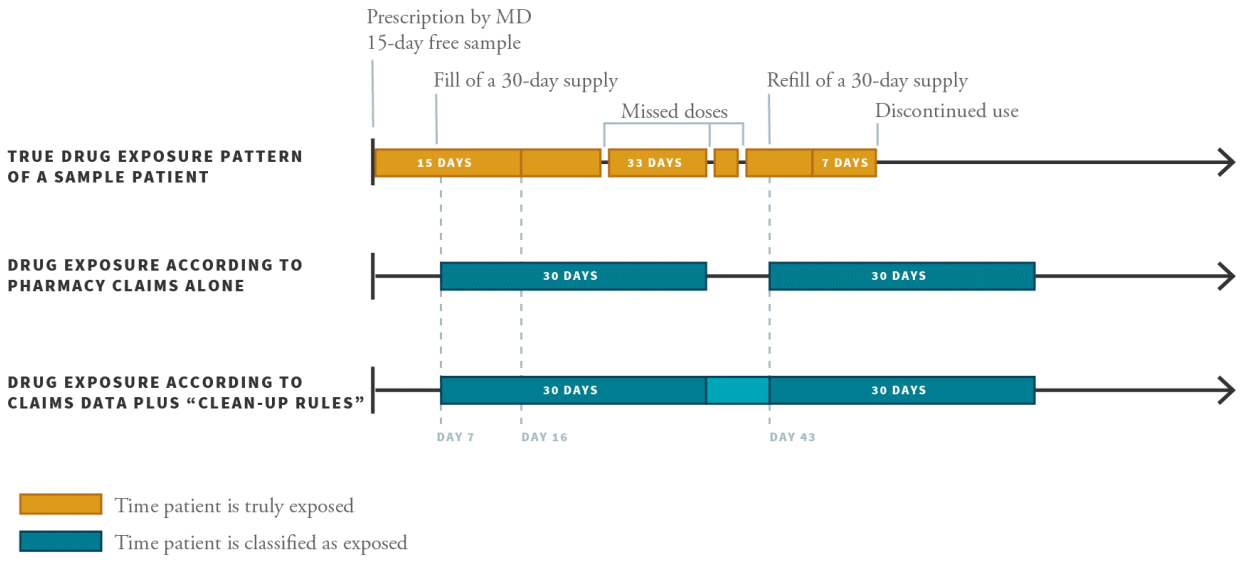
One way to overcome some of these issues is to combine data sources. However, linking records across sources can be technically challenging, and may be inaccurate or incomplete, and may itself introduce some additional biases. Also, there is a need to protect patient privacy.

Research questions require relating exposures to outcomes, which gives us a framework to identify sources of biases and possible errors in the data.

- **Exposure**: Something that could happen to the patient - diseases, drugs, or medical procedures
- **Outcome:** A condition of interest that is assessed as having occurred in the patient timeline, at some point after the exposure

Exposure Misclassified Example:

Suppose a doctor writes a prescription. The prescription is for a 30-day supply of the medication with one refill. The doctor gives the patient a 15-day free sample to cover this gap. The patient starts taking the free sample, and gets the prescription filled at their local pharmacy. Suppose that happens 7 days after the doctor visit. The patient starts the medication from the pharmacy on day 16 after the sample runs out, and continues taking it. Shortly before it runs out, the patient orders a refill, picks that up at the pharmacy, and takes it until the refill runs out. Let's say this sequence of events is what actually happened.



The time the patient is actually exposed to the drug is different from what can be inferred by classifying exposure on the basis of claims data.

We can construct "cleanup rules" that try to produce more accurate estimates by filling in the gaps. A rule might, for example, fill in small gaps, such as less than 7 days, in the record. However, these rules are not perfect.

Outcome Misclassified Example:

A common situation where a patient's outcome could be misclassified is the assignment of codes when the clinical status of a patient is not yet certain.

What can we do to reduce misclassification biases?

- Require multiple mentions of a diagnosis code for us to believe that the patient indeed has the condition
- Require that along with a diagnostic code for the condition of interest, there also must be an occurrence of a disease-specific procedure code
- Require the medication that is very specific to a disease be present in the record to conclude if the patient had that disease

Terminology

- **Electronic phenotyping**: Strategies used to determine if the outcome indeed occurred in the patient's timeline

We can validate the error rate (or misclassification rate) of these strategies for electronic phenotyping by:

- Comparing to a manual chart review of the patient's primary medical record
- Putting bounds on the net-effect of misclassification by constructing a computer simulation that adds random misclassification to the exposure or outcome measurements to see the effect on the answer to the question at hand

When using clinical data, use multiple sources because one source may protect from a weakness in another source or may help in estimating the error rate in another source.

HEALTHCARE DATA SOURCES

Sources of Data:

- Medical record of a patient: Referred to as "charts", stored in computer systems as electronic medical records (EMR) or sometimes also called the electronic health record (EHR)
- Newer kinds of EMR data available in some systems come from genetic tests as well as from consumer devices such as wireless blood pressure cuffs and activity monitors
- Delivery of clinical care produces documentation about the care which includes:
    - Progress notes written by doctors, nurses, and other clinical providers
    - Any orders written by a doctor
    - Results of laboratory tests or imaging studies

EMR Datasets:

- Medical Information Mart for Intensive Care (MIMIC)
- Cerner Health Facts

Claims Data

- Bill: Healthcare professionals collect payment for a medical service. Contains identifying information about the patient, some description of their insurance status, diagnosis and procedure codes as well as requested charges
- Insurance data include everything that insurance paid for, so these records can follow a single patient across multiple providers
- Datasets of Claims:
    - Truven Marketscan Commercial Claims and Encounters
    - Optum Clinformatics Data Mart
    - The Centers for Medicare & Medicaid Services (CMS)

Pharmacies Record:

- The written prescriptions
- When prescriptions were filled
- How the prescriptions were paid for

It is evidence that goes one step beyond the act of writing the prescription. The "drug record" for a single person can be spread out over multiple pharmacy datasets.

Post-marketing surveillance: Governments and other monitoring agencies maintain databases of reported effects and side effects

- Goal: identify serious problems as soon as possible, and then possibly restrict use of the drug or device, or recall it from the market entirely

Sources of **Surveillance Data**:

- The US Food and Drug Administration (FDA)
- FDA Adverse Events Reporting System (FAERS)
- Manufacturer and User Facility Device Experience (MAUDE)
- Centers for Disease Control and Prevention (CDC)
- Registries run by professional societies, governments
    - Examples of societies: the American Board of Family Medicine, the American Society for Clinical Oncology, the American Academy of Ophthalmology
    - Examples of registries: the PRIME registry, CancerLinq, the Intelligent Research in Sight (IRIS) registry
- Registries for particular diseases or conditions:

Population Health Data: Record expenditures by treatment type, medical condition, geographic area.

- Agency for Healthcare Research and Quality (AHRQ)
- National Inpatient Sample (NIS)
- The Medical Expenditure Panel Survey (MEPS)
- The National Health and Nutritional Examination Survey (NHANES)

Patient-generated: Patients can record their own health states and they can choose to provide this information to their doctor, or make the information publicly available for study by others. Patients can report data to centralized databases about their diagnoses, symptoms, treatments, and outcomes.

Examples of patient-generated data sets include:

- Comments made on social networks organized around medical conditions and diseases
- Online portals for patient-reported information about conditions, symptoms and treatments

Researcher-generated: Research is the systematic investigation of biomedical phenomena in order to find valid and generalizable results.

- Example: Researchers can recruit patients into randomized clinical trials in order to systematically study the effectiveness of a treatment. Typically the treatment is compared to either the best existing treatment, or no treatment.
- Clinical trials are usually quite expensive, rigorously analyzed, and their results influence medical practice for a long time. However, by using random assignment to the treatment or the control groups, ***clinical trials are the most reliable source of data to answer questions about causality***.

Clinical trials in the United States have to be registered at www.Clinicaltrials.gov. The *data* from these clinical trials are increasingly becoming available for re-analysis and re-use after the completion of the trial.

There are multiple sources of healthcare data; each source capturing some aspect of what happened to a patient in their care timeline. No one source has the complete picture of the patient timeline, and using multiple sources to answer the same question will lead to more reliable answers.

Questions to ask when considering a particular data source:

1. Is there a well-documented data model?
2. Where are the data from?
3. Are the data accessible?
4. What are the known errors in the data?

Additional questions you should consider:

1. Does this dataset have the data elements corresponding to the patient characteristics that you need to observe?
2. If not, can you use other data elements as a proxy for the characteristics you really want?
3. If you are studying rare conditions, is the dataset large enough to observe those conditions in sufficient numbers?