Stanford MEDICINE

Stanford | Center for Health Education

INTRO TO CLINICAL DATA STUDY GUIDE

MODULE 3 – REPRESENTING TIME, AND TIMING OF EVENTS, FOR CLINICAL DATA MINING

LEARNING OBJECTIVES

1. Explain why timelines are useful for healthcare data
2. Identify the timescales of interest in a clinical research question.
3. Identify which timescales are represented in which kinds of data
4. Explain the difference between explicit and implicit representations of time
5. Describe at least one problem arising from temporal classification of exposures and outcomes
6. Describe non-stationary and why it is important

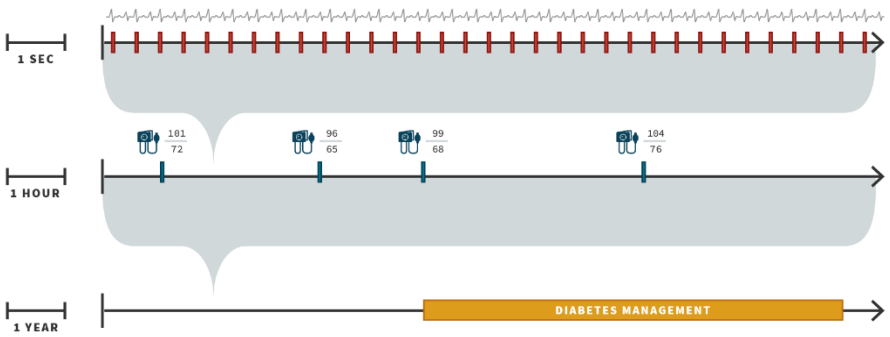
HEALTHCARE HAPPENS OVER TIME

The goal of this module is to discuss how **the patient timeline** relates to **timescales of the questions** we seek to answer as well as the issues that matter when **representing time** and working with healthcare **data that changes over time**.

A good way to integrate all the different sources and types of data for a patient is to place all the events on a timeline. Each patient will have their own timeline. The timeline explicitly captures *when* the patient experienced each event.

We care about **time** for two main reasons.

- A patient's age is a key factor in making medical care decisions about them
- We will often examine the order in which events occur



Copyright © Stanford University

The **timescales** involved in answering medical questions span many orders of magnitude. We may need to consider intervals of well under one second, all the way to a patient's entire lifetime.

The **choice of the timescale** and the way **we choose to represent time** have to deal with the fact that, not only do patients change over time, but the healthcare system itself is evolving over time, so the meaning of the data elements we capture as a by-product of routine care can change without our noticing.

- A **stationary process** is one that generates data that looks similar over time
- A **non-stationary process** is one that generates data whose distribution of values changes over time

In healthcare, the relevant units of time can span many orders of magnitude, from fractions of a second to a century. The relevant interval of time is directly influenced by the disease process, the measurements that current technology can make, and how the healthcare system is organized.

The relevant timescale depends on the question we want to answer. In addition to being influenced by the question, the relevant timescale is also often determined by the kind of data.

Some diseases are chronic, meaning that a cure is not possible, so the disease must be managed indefinitely, perhaps for the rest of the patient's life. Diabetes is an example of a chronic disease.

These two considerations -- **the question**, and **the kind of data** – interact and inform strategy.

Strategy on what features you are going to use:

- How accurately should they be ascertained?
- How many different kinds of features you are going to use?
- How will you infer whether a patient has a condition of interest?

REPRESENTATION OF TIME

How we would capture (or encode) information about time in a computer-based representation -- that is how we *represent time*.

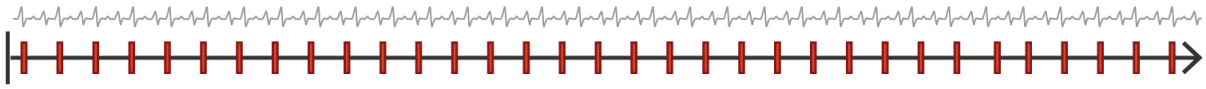
- **Patient-feature matrix**: A rectangular data frame in which each row is a patient and each column is a characteristic about them

In a patient-feature matrix, each row is a patient and each column is a characteristic about them, such as an identifier.

Copyright © Stanford University

A **time series** is a set of measurements that are sampled at regularly-spaced intervals, with each measurement being of the same type.

It is important to know that an important area of healthcare data that uses time series data and related analysis methods is in the ICU. A patient in the ICU is typically extremely ill, and their physiological status needs to be monitored continuously via sensors attached to the patient's body. These streams of data have regular sampling intervals and methods from the field of signal processing can handle these data well.



TIME SERIES DATA

NON-TIME SERIES DATA

Most medical data are not acquired on regular clock ticks. They are sampled asynchronously as determined by necessity. For example, the EKG is a continuous measurement, but blood pressure is measured as needed.

Many medical measurements are acquired in two stages:

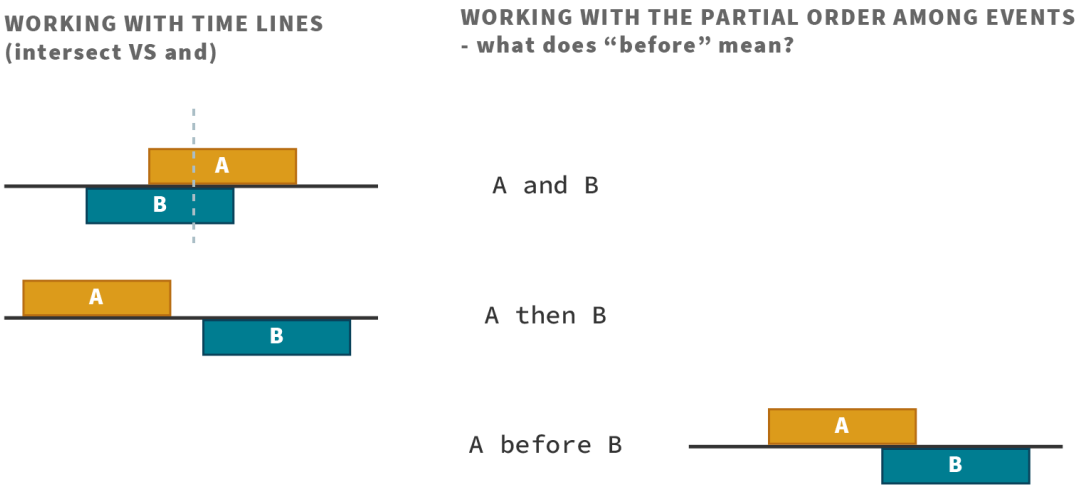
1. The clinician orders the test
2. The test is actually performed

An **indicator variable** marks *what* test was ordered and *when* it was ordered, but does not record the result of the test. The indicator variable is separate from the test value, which records the actual value and the units in which it was measured.

Order of events

Knowing what happened when is a great start, but in many cases, we want to be able to reason about the order of events such as: finding patients with A and B, or patients with A *then* B.

Copyright © Stanford University



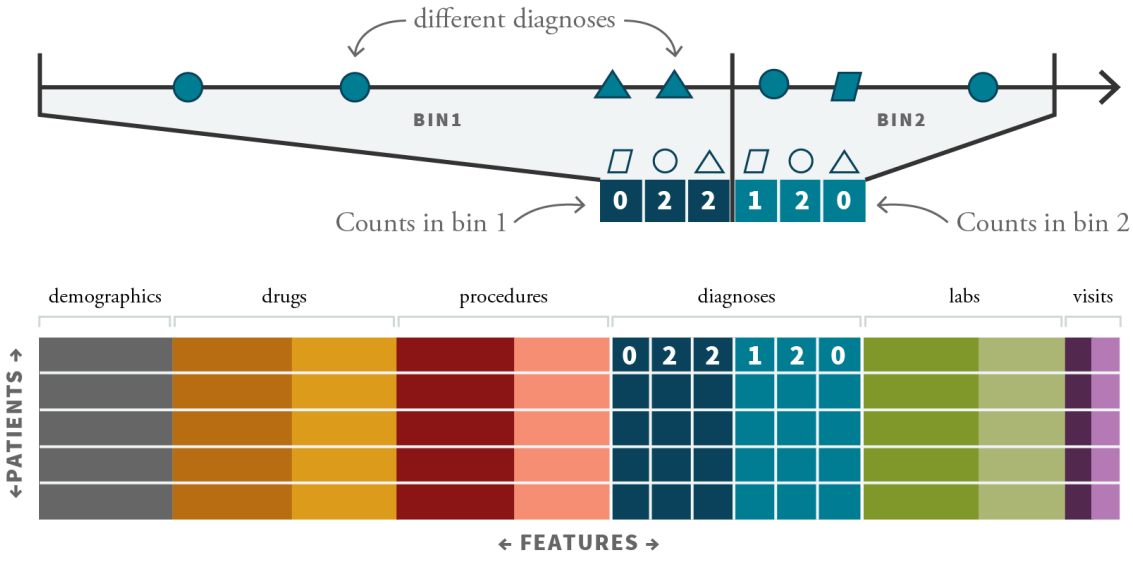
If A and B are instantaneous points in time, then answering this question is straightforward. However, if A and B refer to intervals of having diseases, such as "pneumonia" and "rheumatoid arthritis", then the reasoning becomes more complicated.

What does patients with "pneumonia" and "rheumatoid arthritis" mean? In one interpretation, event A finishes before event B starts. In the second interpretation event A finishes after event B starts.

Most general-purpose databases are not structured to make it easy to work with these distinctions—but a timeline representation does.

Represent time information in the patient-feature matrix

- Binning: Record the number of times that the relevant events occur during specified intervals



Copyright © Stanford University

Let us look at it in detail: We start with a patient timeline. We define time intervals, called bins, that are relevant to the analysis, and count the number of events of each type that occur in each bin. The bin counts become features in the patient-feature matrix.

Choices in time-binning:

- How many bins?
- What granularities of time?
- How does this relate to the timescales at play in your research question?
- How do you aggregate the data within each bin?

Possibilities for making a decision on aggregating or summarizing data within each bin:

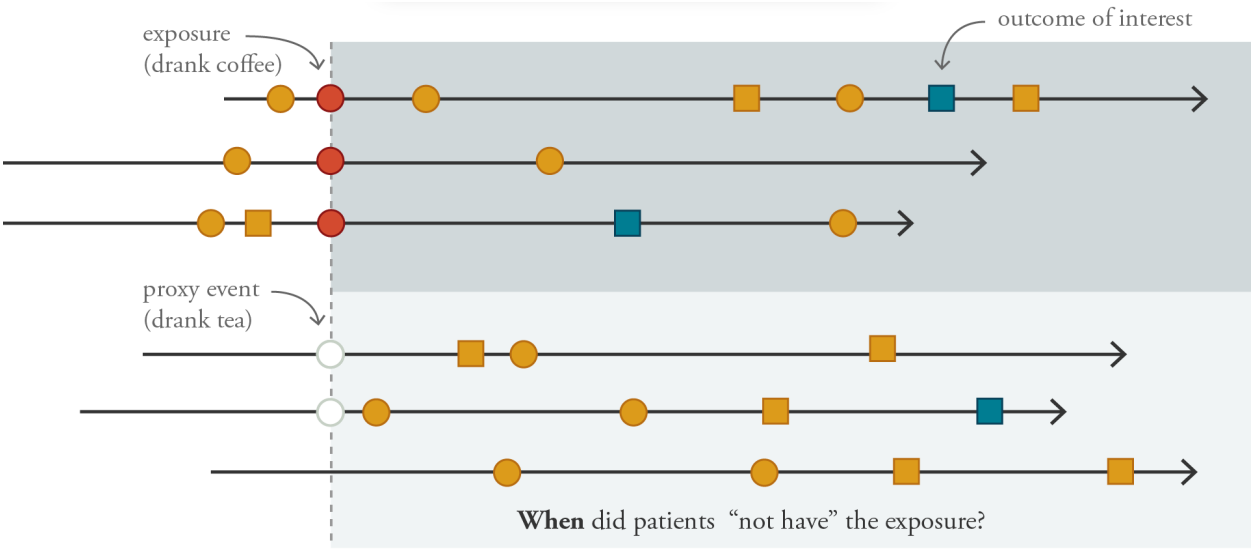
1. It might make sense to use the count of the number of events that occur in each bin as the feature. You could record a count of zero vs a positive count. This would mark absence versus presence.
2. Use the average of all the values in the bin, or the maximum value of all the values in the bin, or the most recent value of all the values in the bin, or the variance of the values in the bin. This choice would be governed by the research question and the clinical item of interest.
3. Add a second feature that records the rate of change of a feature. Creating new features is called feature engineering, and a little medical knowledge goes a long way in crafting such useful features that encode time.

Timing of exposures and outcomes

- **Cohort**: A set of patients that satisfies some inclusion criterion, typically an exposure of some sort
- **Exposure**: Something that could happen to the patient
- **Outcome**: A condition of interest that is assessed as having happened to the patient, usually at some point after the exposure

In general, we are interested in whether there is an association between the exposure and the outcome for patients in the cohort. Therefore, we need to identify those that are exposed and those not exposed as well as those that had the outcome and those who did not. In addition, it matters *when* the exposure and outcome events occurred.

Copyright © Stanford University



For those in the exposed group, such as those who had a coffee, the start time is straightforward to determine. It is when the exposure (i.e. having a coffee) appears on the patient timeline. The start time is also referred to as the index time.

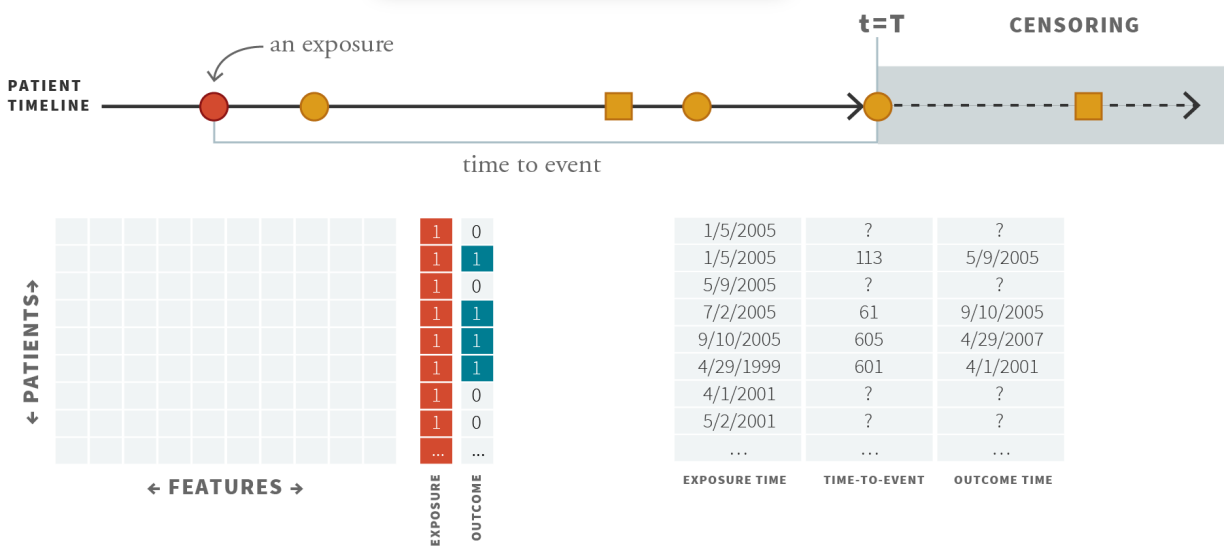
For those non-exposed, meaning those that did not have a coffee, it is straightforward to determine that they are not exposed, because you won't find any exposure or coffee drinking, events on the patient timeline. However, what time should we use for when the non-exposure to coffee happens? This is a very important issue, and not always easy to solve.

A possible solution is to introduce a proxy event (such as drinking tea) whose presence is taken to mark the start of non-exposure. However, this may change who enters the control group and introduces a selection bias. For example, it excludes everyone who did not drink tea or coffee.

Remember: Even if you know *who* is not exposed, it is hard to decide *as of when* they should be counted as not exposed.

We also need the time when the outcome occurs. This will allow us to compute the time-to-event, which is the difference between the time of the outcome and the time of the exposure.

Copyright © Stanford University



Without a time-stamp, we cannot compute the time difference. In epidemiology, this condition is called "right censoring". At a given point in time T, there will be some patients who have not yet developed the outcome.

- One possible solution is to use a special code, which we will write as ">T", which means "the outcome has not happened yet."
- Another solution is to record the time of the outcome, if known, or the time when the patient was last observed along with an indicator variable to mark whether the recorded time is for the outcome or for the last observation.

In summary, you need to diligently determine exposure and outcome times correctly.

DATA CHANGE OVER TIME

Along with the choice of the **timescale** and the **representation of time**, we have to deal with the fact that the healthcare system itself is evolving over time. Collectively, these changes make our **data generation "nonstationary"**.

- **Stationary process**: One that generates data that looks similar over time
- **Non-stationary process**: One that generates data whose distribution of values changes over time

Non-stationarity is usually a problem, but is often ignored.

Copyright © Stanford University

One clever way to use machine learning to *detect* non-stationarity is to remove time as a feature, and then try to predict time from the remaining variables. If we can do that accurately, then there is some pattern in the predictor variables that correlates with time -- meaning there is strong non-stationarity.

The point of the discussion is for you to know that this phenomenon exists, and that you have to test for its presence. This is particularly true for **datasets that span long time intervals** and **studies that have large timescales**.

Copyright © Stanford University