STA130 W12

Monday, December 2, 2019 2:22 PM

Green Tea Study (Kuriyama et al, 2016)

Sample	Study Design
• Population studied restricted to everyone covered by Osaki National Health Insurance • Invited everyone to participate (tried for a CENSUS) • 95% response rate (40530 participants) • Further restricted to healthy adults aged 40-79	• Prospective cohort study • Participants followed up over time, measuring and recording variables on each person

<u>Confounders (i.e. confounding factors or confounding variables)</u>
In an **observational study**, variables are "observed" (measured and recorded) without manipulation of variables or conditions by the researcher
->just observing/recording what is going on
This may lead to **confounding**. Two variables are **confounded** if their effect on the response variable are mixed up (confused) and there is no way to separate them out. If this is the case, we have no way of determining which variable is causing changes to the response.

<u>What are possible confounders for the association between green tea consumption and longevity?</u>
Goal is to determine if drinking tea <u>causes</u> people to live longer

- Hydration in general
- More exercise?
- Socio economic status
- Concern for health
- Stress level?

Because of these(and other) confounders, we can't answer the question above based only on this study.

A 2012 study showed that heavy use of marijuana in adolescence is associated with lower IQ. What are some potential confounders for this association?

- Consuming other drugs
- Skipping school

When we have data from an observational study, we can only conclude **association** between variables, not **causation**.
-> even though we wish we could talk about causation rather than association

<u>Designing studies to avoid confounding</u>
In an **experiment** (or **randomized trial** or **randomized controlled trial**) variables and/or conditions are manipulated by the researcher and the impact on other variable(s) is measured and recorded.
-> eliminating confounders
The key is to randomly assign some individuals to one treatment (or condition) and randomly assign others to another treatment (sometimes this other treatment is a **control**)

> *you can have more than 2 treatments groups too - what is important is that individuals are randomly assigned to them!

The groups (before treatments are applied) should be very similar to each other with respect to the other variables. Any differences between individuals in the treatment and control groups would just be due to random chance!
If there is a **significant difference** in the **outcome** between the two groups, we may have evidence that there is a **causal relationship** between the treatment and the outcome.
-> this works best when the size of the groups is large

<u>Establishing causality</u>
Suppose you were interested in designing a study to determine whether smoking caused lung cancer.
What would a randomized trial look like?

1. Start with a sample
2. Randomly divide sample into 2 groups
   a. Group A: Make them smoke
   b. Group B: Not smoke
3. Follow both groups over time and monitor their health, and whether they get lung cancer

-->> We can't do this...

<u>Causation from Observational Studies?</u>
Although well-designed randomized trials are the best way to establish a causal relationship, observational studies can also help build evidence for causation
Bradford Hill criteria

Strength of association(if variable strong, more likely to be caused)	Plausibility

Consistency	Coherence
Specificity	Experiment
Temporality(cause before effect)	Analogy
Biological gradient	

-> these criteria can't prove causality, but they can help use determine if it is reasonable
-> More criteria is satisficed, more reasonable it is.

The Nuremberg Code

- Ethical codes often emerge out of crisis events
- The Nuremberg code was formulated in August 1947 in Nuremberg, Germany, by American judges sitting in judgement of Nazi doctors accused of conducting murderous and torturous human experiments in concentration camps during the war

The Nuremberg code codified many of our standard principles of ethical research, including:

- research must appropriately balance risk and potential benefits
- researched must be well-versed in their discipline and ground human experiments in animal trials.

Ethical Scandal in the US: Tuskegee Syphilis Study

Study Goal: study the natural progression of syphillis (a disease which can cause serious health problems if le untreated)
Sample:

- In 1932, US government scientists enrolled 400 African American men from Alabama known to be infected with syphillis
- Participants were told the study would last 6 months, but they were actually monitored for over 40 years

Study

- Observational study (Prospective Cohort Study)
- Participants told they were being treated for "bad blood", but received no medical intervention
- Subjects were never told they had syphilis, and denied treatment even after a cure (penicillin) was discovered in 1947

Ethical Scandal: Tuskegee Syphillis Study

By 1972:

- 28 participants had died from syphilis
- 100 had died from conditions related to syphilis
- 40 wives and 19 infants were infected

What was wrong with this study?

- Subjects were not given appropriate information about risks/benefits of the study
- Subjects were not given the opportunity to opt out of the study

On May 16, 1997, President Clinton issued a formal apology for the role of the US government in this study

Ethical Scandals lead to U.S. Law

- Ethical codes did not carry the weight of law in the United States until after a series of scandals in the 1960s and 1970s
- This led to the **1974 National Research Act**, which established the National Commission for the Protection of Human Subjects in Biomedical and Behavioral Research.
- A notable result of the commission was establishing institutional ethics review boards (also known as IRB or REB) which act as independent panels that review research proposals to assess possible harms to human subjects.
- This gives research institutions the power and responsibility to self-regulate through these boards.

Core Principles

- Respect for Persons
- Concern for Welfare
- Justice

Institutional Research Ethics Boards must carefully review research plans in advance to confirm the research abides by these policies.

Free and Informed Consent

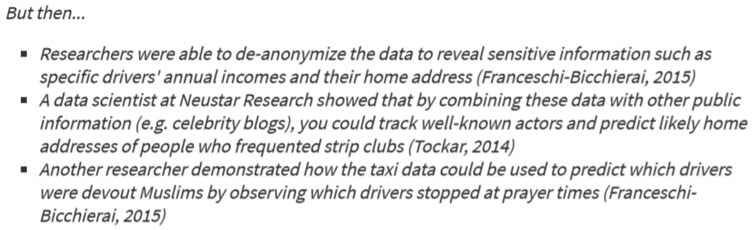
Information: The research procedure, risks and anticipated benefits, alternative procedures (where therapy is involved), and a statement offering the participant the opportunity to ask questions and to withdraw at any time from the research
Comprehension: The manner and context in which information is conveyed is as important as the information itself. For example, presenting information in a disorganized or rapid manner (with too little time to think about it or ask questions), may limit a participant's ability to make an informed choice.
Voluntariness: An agreement to participate in research constitutes a valid consent only if it is voluntary; this requires conditions free of coercion and inappropriate influence.

Ethics in Data Science

Data science involves getting insights about a population by studying existing data, and possibly tying together previously disconnected datasets.
Currently, most ethics boards do not require review for research on existing data, records, and specimens if the data is publicly available
This means that most non-medical data science will receive very little review.

Example: NYC Taxi Data

In 2013, the New York City Taxi and Limousine Commission released a dataset of 173 million individual cab rides.
The dataset included the pickup and dropoff times, locations, fare, and tip amounts
The taxi drivers' medallion numbers were anonymized



But then...

- *Researchers were able to de-anonymize the data to reveal sensitive information such as specific drivers' annual incomes and their home address (Franceschi-Bicchierai, 2015)*
- *A data scientist at Neustar Research showed that by combining these data with other public information (e.g. celebrity blogs), you could track well-known actors and predict likely home addresses of people who frequented strip clubs (Tockar, 2014)*
- *Another researcher demonstrated how the taxi data could be used to predict which drivers were devout Muslims by observing which drivers stopped at prayer times (Franceschi-Bicchierai, 2015)*

Ethics in data science: A few questions to consider

1. Should algorithms be transparent?
2. Are there biases built into algorithms?
3. What does it mean when an algorithm makes the final decision?

Should algorithms be transparent?

Some predictive algorithms give us more than just a prediction: they also give us some insight as to what factor(s) influenced the prediction. Some examples from this course include:

- Linear regression models
- Classification trees (to some extent)

Other algorithms yield predictions, but no information about how it got from the inputs to the prediction

- Neural networks (you may see these in future courses)

What is more important - getting the most accurate predictions, or understanding the factor(s) which influence a prediction?

Example: Identifying effective teachers based on student performance

Goal: Improve education for low-income minority students by collecting data on student performance and using an algorithm to measure teacher performance
Predictor: Student scores on standardized tests
Interpretation: Poor student performance implies ineffective teacher?
Outcome: Some teachers were let go due to their students' poor scores, with no further explanation.
Is this fair? What are potential confounders for the association between a teacher's effectiveness and student scores?

Are there biases built into algorithms?

- Prediction models are taught what they "know" from training data
- Training data can be incomplete, biased, or skewed. This is sometimes referred to as **algorithmic bias**

Example: Amazon recruitment tool

- **Amazon's goal**: predict which applicants they should hire based on their CV in a more unbiased way than humans could
- **Training data**: CV of their current successful employees
- **Is this a good idea?**

-> Hiring people similar to those you already have

Example: Amazon recruitment tool (continued)
Result: Most of the applicants and, therefore, most of the successful applicants, were men in the training data. Therefore, CVs which included the word "women" were penalized by the algorithm.

Is it ethical for Amazon to use this tool to make hiring/promotion decisions?

Remember - Prediction models are only as good as the training data they are built on: if the training data includes/excludes certain groups(for whatever reason), the predictions will reflect this

What does it mean when an algorithm makes the final decision?

Predictive models / algorithms are useful to help us make predictions and gain insights about the world from data. These predictions may have meaningful impacts on people's lives (e.g. teachers getting fired, Amazon's hiring / promotion decisions, etc...)

Another example: Self-driving cars
Over the past several years, significant progress has been made in the development of self-driving cars.
In a self-driving cars, inputs about traffic, road conditions, etc are fed into an algorithm so that the car can adjust in real-time.