

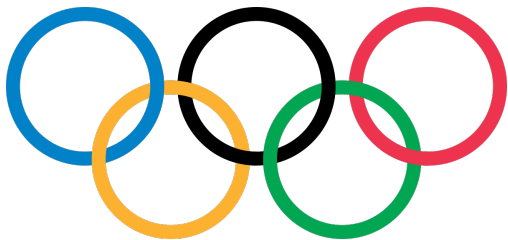
# **Week 12: Ethics, Confounding, and Study Designs**

## **From a Final Exam about Olympics and Cheating**

Scott Schwartz

May 18, 2021

## Olympics, PEDs, and The WADA



**world  
anti-doping  
agency**

Performance enhancing drugs (PEDs) controversy is not uncommon for Olympic athletes. The World Anti-Doping Agency (WADA) – based in Montreal, Quebec – is tasked by the Olympic committee to establish and enforce PED policies; and, “The agency’s key activities include scientific research.”

Routine testing data collected by WADA can be used to study the violation rates with respect to sports and countries. WADA scientists are also ideally placed to study human health and exercise performance. WADA sponsored scientific studies are publicly available through official WADA publications, media outlets interviewing WADA scientists and reporting their results, and in peer-reviewed scientific journals.

## Olympics, PEDs, and The WADA

**What is the most appropriate assessment of the ethical responsibility of WADA with respect to its scientific activities?**

- A. When WADA sponsors and participates in scientific studies, and subsequently engages in efforts to disseminate this information to the public, they are necessarily taking actions with ethical implications and are thus responsible for ensuring that their scientific efforts are carried out in a manner that is consistent with ethical scientific research practices.
- B. Because WADA collects data on publicly observed sports, they are responsible for making it available to the public, and when they release their data (as required) they must be sure to anonymize it by removing the names of athletes so that there is no violation of the re-identification principle.
- C. While the Olympics and WADA impact society through the public activities they organize, they do not have Institutional Review Boards (IRBs) like universities, and so they are not subject to the same ethical considerations and standards as academic research at universities.
- D. The Olympics and WADA are private organizations, and thus their enterprises and activities are not subject to the same ethical standards and considerations that are required of public organizations or publicly funded scientific research projects.

## Morals versus Ethics

Dad. What?

What's the difference between morals and ethics?

Well, a moral's an idea. . . hmm, about right and wrong. And ethics are how those morals apply to specific situations.

Like it's wrong to kill someone, but if they're gonna kill ten people, is it okay to kill them? I don't know why I use this as the example, it's very

Pretty violent

But it's kind of an. . . uh, obvious moral.

This guy Luke, in my book, he has really bad morals and ethics. Both. Huh. How is he immoral and unethical?

'Cause he's really bad. He lies to protect himself, and then he doesn't lie when he could lie to not break Lilly's heart.

I understand that. What do you think he would say to defend his own choices?

Um. . . Uh, well, maybe he feels bad about lying the first time. And so he doesn't lie the second time, even though that would be the right ethic.

Uh-huh. So he would defend his choice by saying that he learned lying was wrong, so now he wants to always tell the truth.

Okay. So he would be right morally but not ethically



## Personal Ethics

Data science gains new insights by tying together existing disconnected data sets

- For publicly available data this generally received very little formal review

### As ethical Statisticians and Data Scientists it is mostly up to us to:

- be accurate in our analyses and conclusions
- be honest in reporting our results even if they're not what we hoped they'd be
- be alert to possible consequences of our results and recommendations
- be respectful of other well-conducted work even if it contradicts our own work
- share credit when our work is based on the ideas of others

Based on the Ethics Guidelines from the [Statistical Society of Canada](#) and the [American Statistical Association](#).

### The Data Science Oath (National Academies of Science, Engineering, and Medicine)

- ① I will not be ashamed to say, "I know not," nor will I fail to call in my colleagues when the skills of another are needed for solving a problem.
- ② I will respect the privacy of my data subjects, for their data are not disclosed to me that the world may know, so I will tread with care in matters of privacy and security.
- ③ I will remember that my data are not just numbers without meaning or context, but represent real people and situations, and that my work may lead to unintended societal consequences, such as inequality, poverty, and disparities due to algorithmic bias.

# Webscrapping

URLs, HTML, CSS, and packages like [polite](#) and [rvest](#) or [Beautiful Soup](#)



BeautifulSoup

---

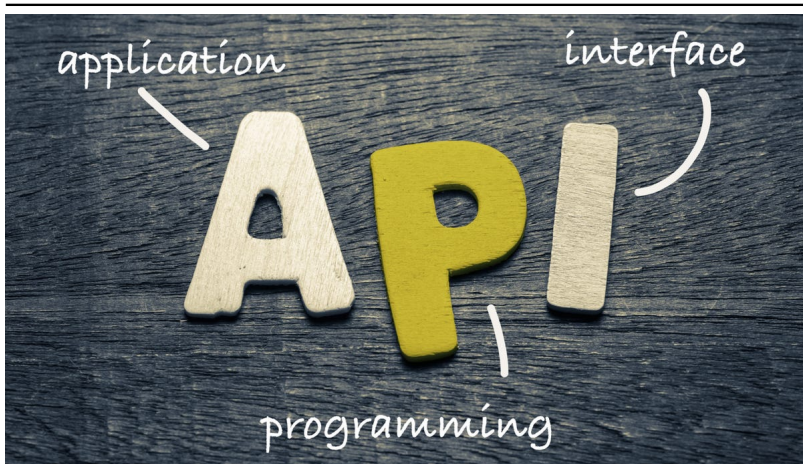
- Websites have *Terms and Conditions* and rules in a file called [robots.txt](#)

I, the [ethical web scraper](#) live by the following principles:

- If you have a public API that provides the data I'm looking for, I'll use it and avoid scraping all together.
- I will always provide a User Agent string that makes my intentions clear and provides a way for you to contact me with questions or concerns.
- I will request data at a reasonable rate. I will strive to never be confused for a DDoS attack.
- I will only save the data I absolutely need from your page. If all I need it OpenGraph meta-data, that's all I'll keep.
- I will respect any content I do keep. I'll never pass it off as my own.
- I will look for ways to return value to you. Maybe I can drive some (real) traffic to your site or credit you in an article or post.
- I will respond in a timely fashion to your outreach and work with you towards a resolution.
- I will scrape for the purpose of creating new value from the data, not to duplicate it.

## APIs

You don't need to know HOW to make crème brûlée to want it and order it. . .



## Societal Ethics

*The Nuremberg code was formulated in August 1947 in Nuremberg, Germany, by American judges judging Nazi doctors accused of conducting murderous and torturous human experiments in concentration camps during the war.*

### The Nuremberg code codified many standard principles for ethical research

such as

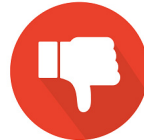
- research must appropriately balance risk and potential benefits
- researchers must be discipline experts and base work on animal trials

*“The voluntary consent of the human subject is absolutely essential. . . The duty and responsibility for ascertaining the quality of the consent rests upon each individual who initiates, directs, or engages in the experiment”*



## Principles of 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:** Agreement to participate in research constitutes valid consent only if it is voluntary based on conditions free of coercion and inappropriate influence.



## Olympics, PEDs, and The WADA

Which of the ethical scientific research principles of free and informed consent, if any, is violated by WADA requiring Olympians to submit blood samples for the purposes of PED testing as a requirement to compete in the Olympics?

---

Compensation principle

Comprehension principle

Information principle

Re-identification principle

Voluntariness principle

None of the principles of  
free and informed consent  
are violated

---

# Protecting Privacy

Consent given for YOU to see information doesn't mean EVERYONE ELSE can too

## Re-identification risks

Information can be accidentally divulged!

2013 NYC Taxi Commission released 173 million trips with anonymized medallion numbers

- Tockar (2014) combined these data with public celebrity blogs to find likely addresses of well-known actors and who frequented strip clubs
- Franceschi-Bicchierai (2015) could STILL identify sensitive information such as drivers' annual incomes, address, and if they followed devout Muslim prayer times
- Metcalfe and Crawford (2016) examine this further

From [Baumer, Kaplan, and Horton \(2021\)](#):

When Massachusetts governor William Weld collapsed in 1996, an MIT graduate student connected this to public release data from the Massachusetts Group Insurance Commission to find Weld's hospitalization records.

The (US) Health Insurance Portability and Accountability Act (HIPAA) was passed that same year defined regulations to ensure that protected health information (PHI) is protected when it is transferred, received, handled, analyzed, or shared by researchers and medical care providers.

# Olympics, PEDs, and The WADA

**Suppose to participate in the Olympics, WADA requires athletes to allow any data collected on them to be used in WADA sponsored scientific research, and**

- because of the dynamic exploratory nature of WADA sponsored scientific research, WADA cannot be specific about exactly how each Olympians data might be used in the future;
- therefore, the agreement with WADA includes the following considerations and stipulations:
  - all research projects must be reviewed and approved by both internal and external independent IRBs (institutional review boards) before they can be carried out to ensure that all research will be done in an ethical manner
  - any time a study is approved all Olympians whose data is being used will be notified, and may inquire about any details of the study they wish, as well as appeal to the IRBs (within a reasonable but limited time frame) if they are concerned that their data is being used unethically
  - because of the potentially sensitive nature of re-identification in this context (e.g., medical information not related to PEDs that Olympians do not wish to be public knowledge), all data will be securely protected and all identities in published data sets will be anonymized using appropriate methods

## Olympics, PEDs, and The WADA

**Do you feel that if WADA makes participation in their research in the manner specified above a requirement for athletes to complete the Olympics, that WADA is behaving ethically as an organization? Why or why not?**

When answering, address two specific ethical research principles of free and informed consent that you feel are the most relevant to your conclusion(s), as well as your rationale for the conclusion you've come to about them.

You are not expected to come a "correct conclusion" in this question; rather, you will be graded based on identifying relevant principles of free and informed consent and considering them with respect to the prompt in a sensible manner that demonstrates that you understand them.

*Since the Olympics are an international activity that strives to bring all the different countries and unique peoples of the world together, we'll assume WADA dedicates sufficient resources to overcoming logistical challenges related to ensuring communication and understanding in all the languages of the world.*

## Olympics, PEDs, and The WADA

To gain support for their work (and help recover from a “recent scandal” regarding what constitutes a PED), WADA feels it should work to improve its profile by more efficiently catching PED violations. Thus, WADA is considering using their (ethically collected) PED testing data to create a model that predicts if a PED test will fail based on the sport and the part of the world in which the testing is being done. WADA plans to use the predictive model that they create to determine where they should focus their PED testing efforts in order to increase the rates at which they catch PED violators.

**How can WADA determine if they expect this predictive model approach to help them to catch more PED violations? I.e., how can WADA validate that their model can indeed actually predict which sports and which parts of the world have increased PED violations?**

*WADA subsequently determined that they expect their algorithmic predictive model approach to help them catch more PED violations, and so they will deploy their predictive model to algorithmically direct PED testing efforts.*

# Algorithmic Bias

## Algorithms are opinions wrapped in code that can make decisions

- What does transparency mean for (possibly proprietary) algorithms?
- What does it mean for an algorithm to have built-in biases?
- What does it mean when an algorithm makes the final decision?

## What would you share about a Decision Tree or a Linear Regression Model?

If poor standardized test scores are predicted for a classroom, should we **fire** the teacher?

## What happens if training data is biased to start with?

**Men historically succeed** in STEM, so algorithms **recommending hiring men** means what?

## What if humans blindly accept and follow algorithmic recommendations?

Could doctor's **increasing use** of AI recommendations be like following GPS **off a cliff**?

## Olympics, PEDs, and The WADA

As the Olympics approach, it begins being reported that while many countries and sports are seeing decreases in their PED violation rates, the overall rates of PED violations have increased because they have been increasing in countries and sports that have historically had higher PED violation rates. When asked to comment on this WADA provided a press release saying

- “Testing resources have been algorithmically directed towards empirical historical patterns of PED cheating which has allowed increased detection of PED violations.”

Subsequently, a reporter from CBC – Canada’s Olympic broadcasting Network – summarized this by saying

- “The countries and sports that cheat are doing it more than ever, while the rest of the world is cheating less. . .”.

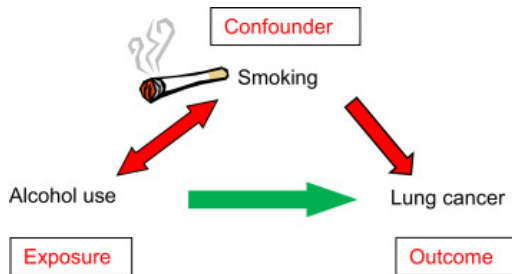
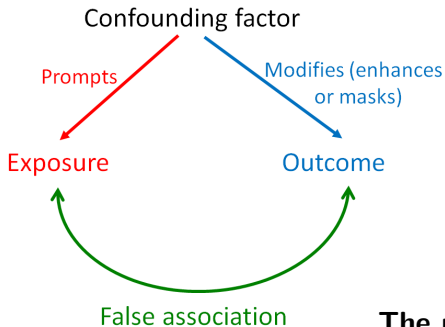
**Do you agree with the CBC reporters statement? Why or why not? Are there any specific biases or risks associated with the strategy WADA is using that you might be concerned about?**



# Confounding

**Multi-collinearity:** observable association prohibits effect attribution

**Confounding:** unobserved variable association *would* prohibit effect attribution

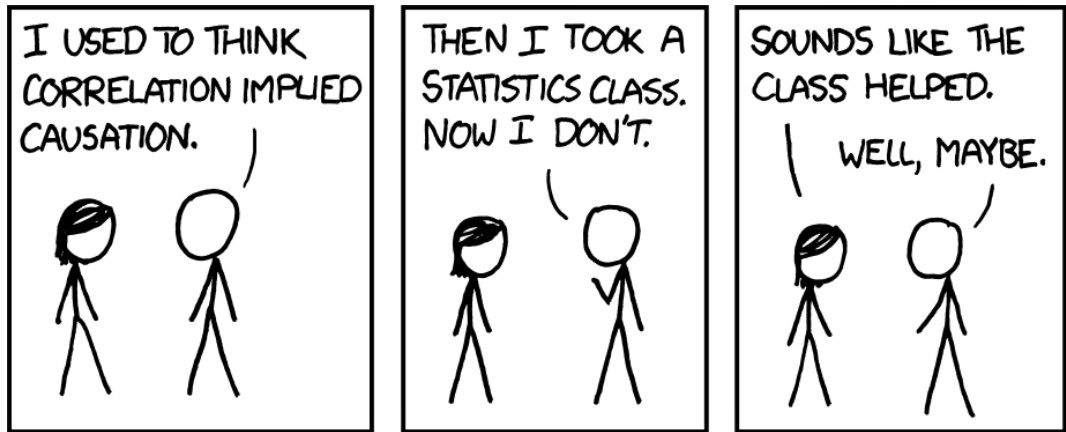


The problem is that you can always just ask...

**Confounding breaks Causality:** Did that "cause" it, or maybe something else?

Opera/Green Tea makes you live longer. Was it from >5 cups/day? Or something else?

## Why Correlation isn't Causation: Confounding



## Olympics, PEDs, and The WADA

A data sleuth reporter from NBC – the official broadcaster of the Olympic games in the US – hopes to increase interest in NBC's Olympic coverage (at CBC's expense) by reporting on how “empirical historical patterns of PED cheating” coincide with increased medal winnings at the Olympics (which will interest Canadians who will be outraged to learn of the unfair play).

- To make the story maximally controversial (to increase viewership even further), after first criticizing WADA for this lack of historical effectiveness, the NBC data sleuth plans to point out the recent WADA critics who note that many of the substances banned and tested for by WADA do not have any known relationship with improved athletic performance. Then the NBC data sleuth will roast WADA by saying that “WADA drug testing is based on policing historical societal norms rather than actually making competition fair”.
- Regardless, the NBC data sleuth reporter supposes that countries that historically violate PED rules more than other countries should also historically have more medals than other countries since it's only now that “algorithms are helping us catch cheaters more effectively”.

Which of the following statements responding to this do you most agree with?

## Olympics, PEDs, and The WADA

- A. Assuming a causal relationship between historical PED prevalence and Olympic medal winnings ignores many possible confounders
- B. Examining associations in observational data violates a cardinal rule of statistical ethics since the analysis could be confounded
- C. Showing that countries with more historical PED violations also have more medals conclusively demonstrates the impact of PEDs
- D. Showing that countries with more historical PED violations don't have more medals is strong evidence that PEDs don't matter

# Study Designs

**Longitudinal** studies take *repeated measurements* at *multiple time points*; whereas, **Cross-Sectional** studies collect a “snapshot” of data only at a *single time point* (although they can ask “historical” questions)

## Observational Study

*Self-Selected Exposure is Observed*

- **Prospective Cohort:**  
outcomes are not yet observed
- **Retrospective Cohort:**  
outcomes have been observed

## Controlled Experiment

*Treatment Assignment is Randomized*

- Random treatments can't be associated with confounders, right?
- The **Placebo Effect** confounds *treatment* and *treatment assignment*

## Case-Control Study

“Good” and “Bad” observed outcomes are paired; then, exposure difference are examined

# Case-Control, Prospective, and Retrospective

## Observational Study Designs: Case Control vs Cohort

Exposure



Disease



Case-Control

Can't use RR, can only use OR because researcher sets the prevalence within the study. Good for rare diseases. In rare diseases, OR approximates RR. In non-rare diseases, the direction of OR and RR are the same, but the actual number obtained of OR and RR are different. You CANNOT obtain a RR for this. It makes no sense to.

Exposure



Disease



Prospective Cohort

RR and OR are both relevant for this. This is sometime used to test out a new intervention/treatment.

Exposure

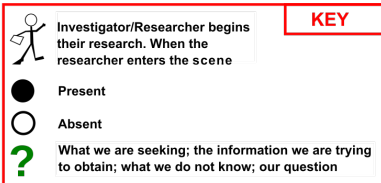


Disease



Retrospective Cohort

RR and OR are both relevant for retrospective cohorts.



- **Case-Control Study:**  
observed outcomes are paired
- **Prospective Cohort Study:**  
outcomes not yet observed
- **Retrospective Cohort:**  
outcomes have been observed

## So observational studies are hopeless?

- Actual “causes” aren’t distinguishable when variables are highly correlated
- In Observational Studies personal actions can be highly correlated
- Measuring something doesn’t mean that you didn’t miss measuring something else
- Randomization in Controlled Experiments can break correlations with treatments
- But Observational Studies can also be used to build evidence...

### Bradford Hill criteria

Strength of association	Consistency
Specificity	Temporality
Biological gradient	Plausibility
Coherence	Experiment
Analogy	

- Hill AB (1965) “The environment and disease: association or causation?”
- Fedak et al (2015) “Applying the Bradford Hill criteria in the 21st century: how data integration has changed causal inference in molecular epidemiology”

## Olympics, PEDs, and The WADA

Suppose WADA has collected the results of its PED testing program on Olympic athletes over the years, and because of growing concerns of long term adverse health effects PEDs, WADA decides it would be ethically irresponsible to not use this data to contribute to public health knowledge and understanding of the long term effects of PEDs. Thus, WADA scientists plan to analyze and report on the long term health outcomes that have been observed in Olympic athletes who have used PEDs.

**From the perspective of appropriate scientific study methodology, what study design component would it be important to have as part of this analysis?**

- A.** A group of Olympic athletes (and outcomes) that could be viewed as a baseline "control group" to be compared against
- B.** A second observational study of people (and outcomes) from different countries and different occupations to compare against
- C.** A randomization component to make this a randomized controlled trial
- D.** A strong public relations effort to make sure that examples of long term adverse health events suffered by athletes who have used PEDs are widely known



## Olympics, PEDs, and The WADA

Continuing the previous example, suppose WADA scientists decided to review long term health outcomes of the Olympic athletes they have tested for PEDs over the years. To do so, they define a group of long term adverse health outcomes and then randomly select Olympic athletes who experienced at least one of these outcomes, and they also randomly select the same number of Olympic athletes who have not experienced any of these outcomes. They then plan to review all the selected Olympians PED testing histories to record if a PED test was ever failed by an athlete.

**What kind of study design is this?**

- A. Case-control study
- B. Prospective cohort study
- C. Randomized controlled trial
- D. Retrospective cohort study