Causal Inference: prediction, explanation, and intervention

Lecture 1: Introduction

Samantha Kleinberg
samantha.kleinberg@stevens.edu
About me: Samantha Kleinberg

- PhD in CS
- Have worked in biomathematics, bioinformatics/computational biology, biomedical informatics
- Current research
  - Time series data, causal inference
  - Simulation
  - Mobile health
“Most striking, society will need to shed some of its obsession for causality in exchange for simple correlations: not knowing why but only what.”

Causal claims abound

Sleep Apnea Tied to Increased Cancer Risk
By ANAHAD O’CONNOR

Two new studies have found that people with sleep apnea, a common disorder that causes snoring, fatigue and dangerous pauses in breathing at night, have a higher risk of cancer. The new research marks the first time that sleep apnea has been linked to cancer in humans.

About 28 million Americans have some form of sleep apnea, though many cases go undiagnosed. For sleep doctors, the condition is a top concern because it deprives the body of oxygen at night and often coincides with cardiovascular disease, obesity and diabetes.

“This is really big news,” said Dr. Joseph Golish, a professor of sleep medicine with the MetroHealth System in Cleveland who was not involved in the research. “It’s the first time this has been shown, and it looks like a very solid association,” he said.

Dr. Golish, the former chief of sleep medicine at the Cleveland Clinic, said that the cancer link may not prove to be as strong as the well-documented relationship between sleep apnea and cardiovascular disease, “but until
Causal claims abound

Sleep Apnea Tied to I Risk

By ANAHAD O'CONNOR

Two new studies have found with sleep apnea, a common causes snoring, fatigue and pauses in breathing at night risk of cancer. The new rese first time that sleep apnea h to cancer in humans.

About 28 million Americans form of sleep apnea, though undiagnosed. For sleep doc condition is a top concern b deprives the body of oxyg often coincides with cardiov disease, obesity and diabete

“This is really big news,” sai Golish, a professor of sleep the MetroHealth System in “It’s the first time this has b association,” he said.

Dr. Golish, the former chief that the cancer link may not relationship between sleep apnea and cancer in the results have been inconclusive.

May 17, 2012 11:24 AM

Two cups of coffee a day cuts risk of dying by 10 percent, research shows

By CBS News Staff

(CBS/AP) How good is coffee for your health? For years, research has gone both ways, with some studies finding it boosts risk for heart disease, while other studies find it could be protective against breast and skin cancers.

Green coffee beans may lead to weight loss, study shows Coffee helps prevent skin cancer? What study shows PICTURES: Coffee and your health: Latest findings

A large-scale study of 400,000 people offers good news for coffee-drinkers: you might just live longer.

The study is the largest ever done on the issue, and the results should reassure any coffee lovers who think it’s a guilty pleasure that may do harm. And whether it’s regular or decaf doesn’t even matter.

"There may actually be a modest benefit of coffee drinking," said lead researcher Neal Freedman of the National Cancer Institute.

The study, published online in the May 16 issue of the New England Journal of Medicine, kicked off in 1995 and involved 402,260 AARP members ages 50 to 71 who lived in California, Florida, Louisiana, New Jersey, North Carolina, Pennsylvania and Atlanta and Detroit. People who already
Causal claims abound

March 12, 2012

**Risks: More Red Meat, More Mortality**

By NICHOLAS BAKALAR

Eating red meat is associated with a sharply increased risk of death from cancer and heart disease, according to a new study, and the more of it you eat, the greater the risk.

The analysis, published online Monday in Archives of Internal Medicine, used data from two studies that involved 121,342 men and women who filled out questionnaires about health and diet from 1980 through 2006. There were 23,926 deaths in the group, including 5,910 from cardiovascular disease and 9,464 from cancer.

People who ate more red meat were less physically active and more likely to smoke and had a higher body mass index, researchers found. Still, after controlling for those and other variables, they found that each daily increase of three ounces of red meat was associated with a 12 percent greater risk of dying over all, including a 16 percent greater risk of cardiovascular death and a 10 percent greater risk of cancer death.

The increased risks linked to processed meat, like bacon, were even greater: 20 percent over all, 21 percent for cardiovascular disease and 16 percent for cancer.

If people in the study had eaten half as much meat, the researchers estimated, deaths in the group would have declined 9.3 percent in men and 7.6 percent in women.

Previous studies have linked red meat consumption and mortality, but the new results suggest a surprisingly strong link.

“When you have these numbers in front of you, it’s pretty staggering,” said the study’s lead author, Dr. Frank B. Hu, a professor of medicine at Harvard.
Correlation = 0!

Correlation = 0!

Now introduce yourself
What is a cause?
The case of Ronald Opus

- He jumped from a 10-story building, planning to commit suicide (there was a note)

- Around the 9th floor, he was hit by the bullet

- He didn’t know that there was a safety net on the 8th floor

- Because of the net, the suicide would have failed
The story continues

• Across the street was an old man and his wife

• He threatened her with the gun, but lost control and the bullet went through the window

• According to husband and wife, neither knew gun was loaded

• Was it an accident?
The plot thickens

- Their son knew about their arguments and loaded the gun, as he was upset with mother

- It turns out, he was so upset that his plan didn’t work, that he jumped out the window, only to be killed by a shotgun on his way down
Causality has consequences

• Sally Clark’s 1\textsuperscript{st} son died in 1996, as a result of SIDS
• Her 2\textsuperscript{nd} son died in 1999, also as a result of SIDS

• Prosecutors argued too unlikely to both be SIDS, must be murder.
  • Chance of SIDS = 1/8,543 so chance of 2 deaths = 1/(8,543*8,543) (approx 1 in 73 million)
  • What’s wrong with this?
Why is causality hard?

• No single definition
• No fail-proof method for finding it
• Observational data
Administrativia

- Course website: http://www.cs.stevens.edu/~skleinbe/teaching/CI16/
- Prerequisites: none
- Textbook: none, book chapter + articles
- Topics/readings on syllabus MAY CHANGE (will make announcement in class)
- Workload/grading:
  - midterm exam (30%), final project (50%), participation (15%), homework (5%)
  - Doing discussion readings is critical!
Some previous final projects

• Does popularity cause campaign contributions?
• What causes flight delays at Newark airport?
• Is the value of bitcoin driven by exchange rates?
• Does fracking cause earthquakes?
Detection of an Infectious Retrovirus, XMRV, in Blood Cells of Patients with Chronic Fatigue Syndrome


Chronic fatigue syndrome (CFS) is a debilitating disease of unknown etiology that is estimated to affect 17 million people worldwide. Studying peripheral blood mononuclear cells (PBMCs) from CFS patients, we identified DNA from a human gammaretrovirus, xenotropic murine leukemia virus–related virus (XMRV), in 68 of 101 patients (67%) as compared to 8 of 218 (3.7%) healthy controls. Cell culture experiments revealed that patient-derived XMRV is infectious and that both cell-associated and cell-free transmission of the virus are possible. Secondary viral infections were established in uninfected primary lymphocytes and indicator cell lines after their exposure to activated PBMCs, B cells, T cells, or plasma derived from CFS patients. These findings raise the possibility that XMRV may be a contributing factor in the pathogenesis of CFS.

XMRV leads to CFS?
XMRV leads to CFS?

Study Links Chronic Fatigue to Virus Class
By DAVID TULLER

August 23, 2010

When the journal Science published an attention-grabbing study last fall linking chronic fatigue syndrome to a recently discovered retrovirus, many experts remained skeptical — especially after four other studies found no such association.

Now a second research team has reported a link between the fatigue syndrome and the same class of virus, a category known as MLV-related viruses. In a paper published Monday by The Proceedings of the National Academy of Sciences, scientists found gene sequences from several MLV-related viruses in blood cells from 32 out of 37 chronic-fatigue patients but only 3 of 44 healthy ones.

The researchers did not find XMRV, the specific retrovirus identified in patients last fall. But by confirming the presence of a cluster of genetically similar viruses, the new study represents a significant advance, experts and advocates say.

“I think it settles the issue of whether the initial report was real or not,” said K. Kimberly McCleary, president of the CFIDS Association of America, the leading organization for people with chronic fatigue syndrome.

Leonard A. Jason, a professor of psychology at DePaul University and a leading researcher on the syndrome, agreed. “This class of retroviruses is probably going to be an important piece of the puzzle,” he said.

Chronic fatigue syndrome, estimated to afflict at least one million Americans, has no known cause and no accepted diagnostic tests, although patients show signs of immunological, neurological and endocrinological abnormalities. Besides profound exhaustion, symptoms include sleep disorders, cognitive problems, muscle and joint pain, sore throat and headaches.

The new paper, by researchers from the National Institutes of Health, the Food and Drug Administration and Harvard Medical School, was accepted for publication in May. Social networks and online communities soon learned the general findings and were eagerly awaiting the paper.

But in July, researchers from another federal agency, the Centers for Disease Control and Prevention, published a report finding no such retrovirus in blood cells from chronic fatigue patients.
But ...

RETRACTION

Post date 23 December 2011

Science is fully retracting the Report “Detection of an infectious retrovirus, XMRV, in blood cells of patients with chronic fatigue syndrome” (1). Multiple laboratories, including those of the original authors (2), have failed to reliably detect xenotropic murine leukemia virus–related virus (XMRV) or other murine leukemia virus (MLV)–related viruses in chronic fatigue syndrome (CFS) patients. In addition, there is evidence of poor quality control in a number of specific experiments in the Report. Fig. 1, table S1, and fig. S2 have been retracted by the authors (3). In response to concerns expressed about Fig. 2C (summarized in (4)), the authors acknowledged to Science that they omitted important information from the legend of this figure panel. Specifically, they failed to indicate that the CFS patient–derived peripheral blood mononuclear cells (PBMCs) shown in Fig. 2C had been treated with azacytidine as well as phytohemagglutinin and interleukin-2. This was in contrast to the CFS samples shown in Figs. 2A and 2B, which had not been treated with azacytidine.

Given all of these issues, Science has lost confidence in the Report and the validity of its conclusions. We note that the majority of the authors have agreed in principle to retract the Report but they have been unable to agree on the wording of their statement. It is Science’s opinion that a retraction signed by all the authors is unlikely to be forthcoming. We are therefore editorially retracting the Report. We regret the time and resources that the scientific community has devoted to unsuccessful attempts to replicate these results.

BRUCE ALBERTS
Editor-in-Chief

References
Science is fully retracting the Report “Detection of an infectious retrovirus, XMRV, in blood cells of patients with chronic fatigue syndrome” (1). Multiple laboratories, including those of the original authors (2), have failed to reliably detect xenotropic murine leukemia virus–related virus (XMRV) or other murine leukemia virus (MLV)–related viruses in chronic fatigue syndrome (CFS) patients. In addition, there is evidence of poor quality control in a number of specific experiments in the Report. Fig. 1, table S1, and fig. S2 have been retracted by the authors (3). In response to concerns expressed about Fig. 2C [summarized in (4)], the authors acknowledged to Science that they omitted important information from the legend of this figure panel. Specifically, they failed to indicate that the CFS patient–derived peripheral blood mononuclear cells (PBMCs) shown in Fig. 2C had been treated with azacytidine as well as phytohemagglutin and interleukin-2. This was in contrast to the CFS samples shown in Figs. 2A and 2B, which had not been treated with azacytidine.

Given all of these issues, Science has lost confidence in the Report and the validity of its conclusions. We note that the majority of the authors have agreed in principle to retract the Report but they have been unable to agree on the wording of their statement. It is Science’s opinion that a retraction signed by all the authors is unlikely to be forthcoming. We are therefore editorially retracting the Report. We regret the time and resources that the scientific community has devoted to unsuccessful attempts to replicate these results.

References
Why do we need causes?

- Prediction
- Explanation
- Intervention
Google flu

Detecting influenza epidemics using search engine query data

Jeremy Ginsberg¹, Matthew H. Mohebbi¹, Rajan S. Patel¹, Lynnette Brammer², Mark S. Smolinski¹ & Larry Brilliant¹

¹Google Inc. ²Centers for Disease Control and Prevention
The Parable of Google Flu: Traps in Big Data Analysis

David Lazer, Ryan Kennedy, Gary King, Alessandro Vespignani
Prediction

Blackouts → match sales
Season → match sales
Smoking rate → lung cancer rate
Prediction, continued

Gene mutation

- lower exercise tolerance
- Disease A
I used to think correlation implied causation.

Then I took a statistics class. Now I don't.

Sounds like the class helped. Well, maybe.

http://xkcd.com/552/
Correlation

(a) Positively correlated

(b) Uncorrelated
Correlations abound

- High HDL is related to lower heart disease
- Height and age
- Tides and traffic on the west side highway
- XMRV and CFS
Correlation and Causation

• What’s a correlation?
  • Relatedness of variables across samples or time

• Common measure: Pearson’s correlation coefficient

\[ r = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}} \]
Measuring correlation

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

\[
r = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2} \sqrt{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}}
\]

Cups of coffee (X) vs. correct text answers (Y)

\[
r = \frac{(5.2 + 3.2 + 0 + 2.8 + 4.8)}{4.1473 \times 4}
\]

\[
r = 0.9645
\]
(a) Strong positive ($r=0.963$)

(b) Strong negative ($r=-0.963$)
(a) Strong positive \( r=0.963 \)

(b) Strong negative \( r=-0.963 \)

(c) Weaker positive \( r=0.788 \)

(d) Weaker negative \( r=-0.788 \)
(a) Strong positive ($r=0.963$)

(b) Strong negative ($r=-0.963$)
(e) No correlation \((r=0.000)\)

(f) Nonlinear relationship \((r=0.000)\)
Hidden Common Causes
Hidden Common Causes
Correlation with some causation
Nonstationary time series

http://bama.ua.edu/~sprentic
Nonstationary time series

- Starbucks stores
- Autism diagnoses

Restricted range
Canceling out
Simpson’s paradox

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Dead</th>
<th>Alive</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>85</td>
<td>215 (72%)</td>
</tr>
<tr>
<td>B</td>
<td>59</td>
<td>241 (80%)</td>
</tr>
<tr>
<td>Total</td>
<td>144</td>
<td>456</td>
</tr>
</tbody>
</table>
Causation without correlation: Simpson’s paradox

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Men</th>
<th></th>
<th>Women</th>
<th></th>
<th>Combined</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dead</td>
<td>Alive</td>
<td>Dead</td>
<td>Alive</td>
<td>Dead</td>
<td>Alive</td>
</tr>
<tr>
<td>A</td>
<td>80</td>
<td>120 (60%)</td>
<td>5</td>
<td>95 (95%)</td>
<td>85</td>
<td>215 (72%)</td>
</tr>
<tr>
<td>B</td>
<td>20</td>
<td>20 (50%)</td>
<td>39</td>
<td>221 (85%)</td>
<td>59</td>
<td>241 (80%)</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>140</td>
<td>44</td>
<td>316</td>
<td>144</td>
<td>456</td>
</tr>
</tbody>
</table>

Explanation (1)

- Why are two variables related?

Diagram:
- Diabetes
  - Blurred vision
  - Weight loss
- CKD
  - Renal failure
  - Medication
Explanation (2)

- General causes of illness vs. cause of a specific patient’s illness

- Why did an event happen?
  - Why did a particular person develop lung cancer at age 42?
  - What led to the U.S. recession in 2007?
  - Is a stroke patient’s secondary brain injury due to seizures?
Automating explanation

- Methods for finding causes from data, but what about explaining events?
- Practical problem, but challenging
  - Information incomplete
  - Where do explanations come from?
  - General and singular can differ
Intervention

• Why do we need causes to take action?
  • Buying stocks
  • Taking vitamins
  • Decreasing sodium to prevent hypertension

• What happens if we intervene on a correlated factor?
Using causes to guide intervention

Blackouts → match sales
Season → match sales
Smoking rate → lung cancer rate
Using interventions to find causes

- Does playing violent video games make children violent?
- Does too little sleep increase mortality rate?
- Does medication cause side effects?
Recap

A cause
....allows prediction of future events
....explains connections
....explains occurrences
....enables interventions to prevent/produce outcomes

BUT! Not every cause does and the story is more complicated
Causality & time
Interpretation in the absence of time

• Compare:

• A. Smoking causes lung cancer with probability $\approx 1$ after 90 years

• B. Smoking causes lung cancer with probability $= \frac{1}{2}$ in less than 10 years.
Probability

- Few relationships deterministic
  - Relationship vs. limits of knowledge
- Understanding risk
- Choosing intervention target
  - Medication efficacy vs. side effects
- Can also measure strength of relationship
Complexity

- Interactions
  - Smoking + lung cancer: other conditions affecting probability and time: genetics, environment

- Planning effective interventions
  - Political Speeches

- Durations, conjunctions, sequences of events
Does coffee prevent or hasten death?
Course overview

Weeks 1-3
What Is a cause?

Weeks 4-8
How can we find causes?

Weeks 9-12
When can we find causes?

Weeks 13-14
Projects + Special topics
Three main questions

• What is a cause?
  • Theories of what distinguishes them from correlations and how we can identify them

• How can we find causes?
  • Features of causes that allow us to learn about them

• When can we infer causes?
  • Methods for inference from data
  • Study design
  • Applications to challenging cases
• **Causal inference**: finding causal relationships from data

• **Causal explanation**: finding reason for a specific event that occurs at a particular time and place
Mo data mo problems

• Big ≠ good
• Uncertainty
• Selection bias
• Signal:noise
• Interpretation
• Time
• Ground truth
Many unsolved problems

- Hidden variables
- Relationships that change over time
- Hypothesis generation
- Estimating uncertainty
- Testing assumptions
- Automating explanation
Philosophy:
What is a cause?
Philosophy:
What is a cause?

Computer Science:
How can we automate inference/explanation?

How do we learn of causes?
Psychology:
How do we gain and use causal knowledge?

Philosophy:
What is a cause?

Computer Science:
How can we automate inference/explanation?

What's the relationship between moral and causal judgment?

How do we learn of causes?
Psychology:
How do we gain and use causal knowledge?

Philosophy:
What is a cause?

Computer Science:
How can we automate inference/explanation?

Epidemiology:
What affects human health?

Large-scale analysis of EHRs

What's the relationship between moral and causal judgment?

How do we learn of causes?
Psychology: How do we gain and use causal knowledge?

Philosophy: What is a cause?

Computer Science: How can we automate inference/explanation?

Medicine/biology: Applications to neuroscience, genomics

Epidemiology: What affects human health?

What's the relationship between moral and causal judgment?

How do we learn of causes?

BNs

RCTs

Large-scale analysis of EHRs
Psychology:
- How do we gain and use causal knowledge?

Philosophy:
- What is a cause?
- What's the relationship between moral and causal judgment?

Economics:
- Do policies achieve goals?

Computer Science:
- How can we automate inference/explanation?

Medicine/biology:
- Applications to neuroscience, genomics

Epidemiology:
- What affects human health?
- Large-scale analysis of EHRs

Granger causality
Granger causality outside economics

Economics:
Do policies achieve goals?

Medicine/biology:
Applications to neuroscience, genomics

Granger causality outside economics

**Economics:**
Do policies achieve goals?

**Medicine/biology:**
Applications to neuroscience, genomics


EHR analysis

Computer Science:
How can we automate inference/explanation?

Epidemiology:
What affects human health?

Large-scale analysis of EHRs

Bar chart showing:
- dx_hypothyroidism: 21 cases
- dx_diabetes: 22 cases
- dx_overweight: 22 cases
- first_antihtn_combo: 10 cases
- dx_urinary_symptoms: 11 cases
Computational biology

**Medicine/biology:**
Applications to neuroscience, genomics

**Computer Science:**
How can we automate inference/explanation?

---

Causal judgment

Psychology:
How do we gain and use causal knowledge?

Philosophy:
What is a cause?

What's the relationship between moral and causal judgment?

The receptionist in the philosophy department keeps her desk stocked with pens. The administrative assistants are allowed to take the pens, but faculty members are supposed to buy their own.

The administrative assistants typically do take the pens. Unfortunately, so do the faculty members. The receptionist has repeatedly emailed them reminders that only administrative assistants are allowed to take the pens.

On Monday morning, one of the administrative assistants encounters Professor Smith walking past the receptionist’s desk. Both take pens. Later that day, the receptionist needs to take an important message... but she has a problem. There are no pens left on her desk.

-Professor caused it?
-Assistant caused it?
Causal judgment

The receptionist in the philosophy department keeps her desk stocked with pens. The administrative assistants are allowed to take the pens, but faculty members are supposed to buy their own.

The administrative assistants typically do take the pens. Unfortunately, so do the faculty members. The receptionist has repeatedly emailed them reminders that only administrative assistants are allowed to take the pens.

On Monday morning, one of the administrative assistants encounters Professor Smith walking past the receptionist’s desk. Both take pens. Later that day, the receptionist needs to take an important message… but she has a problem. There are no pens left on her desk.

-Professor caused it?
-Assistant caused it?

18 students, -3 to 3 scale
Professor: 2.2
Assistant: -1.2
Finance

Is the market efficient?
## Finance

What drives stock prices?

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>FDR</th>
<th>FNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuous</td>
<td>0.003</td>
<td>0.048</td>
</tr>
<tr>
<td>Discrete</td>
<td>0.010</td>
<td>0.048</td>
</tr>
</tbody>
</table>

[Graph showing distribution of z-values with different color lines representing data, f(z), empirical null, and theoretical null.]

[Graph showing a time series with what appears to be stock price fluctuations.]
Politics

- Do speeches affect presidential popularity ratings?
- Bush’s speech key-phrases versus polling numbers
Turing award + Nobel prize
For next week

Rasputin was led to a cellar and served cake and wine laced with cyanide – enough poison to kill 5 men.
For next week

Rasputin was led to a cellar and served cake and wine laced with cyanide – enough poison to kill 5 men.
This failed, so he was shot in the back.
For next week

Rasputin was led to a cellar and served cake and wine laced with cyanide – enough poison to kill 5 men.

This failed, so he was shot in the back.

Once again Rasputin survived…only to be shot again.
For next week

Rasputin was led to a cellar and served cake and wine laced with cyanide – enough poison to kill 5 men.
This failed, so he was shot in the back.
Once again Rasputin survived…only to be shot again. They then bound him and threw him in an icy river.
Rasputin was led to a cellar and served cake and wine laced with cyanide – enough poison to kill 5 men.
This failed, so he was shot in the back.
Once again Rasputin survived…only to be shot again.
They then bound him and threw him in an icy river.
He escaped the bonds, but drowned.
For next week

Rasputin was led to a cellar and served cake and wine laced with cyanide – enough poison to kill 5 men.
This failed, so he was shot in the back.
Once again Rasputin survived…only to be shot again.
They then bound him and threw him in an icy river.
He escaped the bonds, but drowned.
What caused his death?
For next week

Rasputin was led to a cellar and served cake and wine laced with cyanide – enough poison to kill 5 men.
This failed, so he was shot in the back. Once again Rasputin survived…only to be shot again. They then bound him and threw him in an icy river. He escaped the bonds, but drowned. What caused his death?

See syllabus for reading