Predicting and evaluating how changes in exposures change health risks

Tony Cox

April 29, 2018
Goals

• *Predict* changes in public health effects caused by changes in exposures
  – Not associations or slopes, but changes over time
• *Evaluate* changes in effects caused by changes in exposures in hindsight (accountability)
  – Model data on changes, not just levels
• Use trustworthy methods, get objective answers
  – Do not rely on untested assumptions or counterfactual comparisons (Dublin)
  – Use automated algorithms to avoid p-hacking
  – Discover who benefits, how, and how much from reduced exposures to air pollution
Causal questions

- Statistical inference question:
  - How does the conditional probability distribution for observed daily death count (AllCause75) depend on observed values of other variables?
    - $P(\text{deaths} \mid \text{tmin, PM2.5, etc.})$

- Causal question:
  - How does the conditional probability distribution for observed daily death count (AllCause75) change in response to changes in values of other variables?
    - $P(\text{deaths} \mid \text{tmin, do(PM2.5), etc.})$
    - How would exogenously reducing PM2.5, tmax, etc. change elderly mortality, AllCause75?

- Seeing ≠ doing! (Pearl, 2009)

- This talk: Illustrate machine learning (ML) techniques for predicting causal impacts with minimal assumptions
Causal questions

- Statistical inference question:
  - How does the conditional probability distribution for observed daily death count (AllCause75) depend on observed values of other variables?
    - $P(\text{deaths} \mid t\text{min}, \text{PM2.5}, \text{etc.})$
- Causal question:
  - How does the conditional probability distribution for observed daily death count (AllCause75) change in response to changes in values of other variables?
    - $P(\text{deaths} \mid t\text{min}, \text{do(PM2.5)}, \text{etc.})$
  - How would exogenously reducing PM2.5, $t\text{max}$, etc. change elderly mortality, AllCause75?
- Seeing ≠ doing! (Pearl, 2009)
- This talk: Illustrate machine learning (ML) techniques for predicting causal impacts with minimal assumptions

<table>
<thead>
<tr>
<th>year</th>
<th>month</th>
<th>day</th>
<th>AllCause75</th>
<th>PM2.5</th>
<th>t\text{min}</th>
<th>t\text{max}</th>
<th>MAXR\text{H}</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007</td>
<td>1</td>
<td>1</td>
<td>151</td>
<td>38.4</td>
<td>36</td>
<td>72</td>
<td>68.8</td>
</tr>
<tr>
<td>2007</td>
<td>1</td>
<td>2</td>
<td>158</td>
<td>17.4</td>
<td>35</td>
<td>75</td>
<td>48.9</td>
</tr>
<tr>
<td>2007</td>
<td>1</td>
<td>3</td>
<td>139</td>
<td>19.9</td>
<td>44</td>
<td>75</td>
<td>61.3</td>
</tr>
<tr>
<td>2007</td>
<td>1</td>
<td>4</td>
<td>164</td>
<td>64.6</td>
<td>37</td>
<td>68</td>
<td>87.9</td>
</tr>
<tr>
<td>2007</td>
<td>1</td>
<td>5</td>
<td>136</td>
<td>6.1</td>
<td>40</td>
<td>61</td>
<td>47.5</td>
</tr>
<tr>
<td>2007</td>
<td>1</td>
<td>6</td>
<td>152</td>
<td>18.8</td>
<td>39</td>
<td>69</td>
<td>39</td>
</tr>
<tr>
<td>2007</td>
<td>1</td>
<td>7</td>
<td>160</td>
<td>19.1</td>
<td>39</td>
<td>69</td>
<td>39</td>
</tr>
<tr>
<td>2007</td>
<td>1</td>
<td>8</td>
<td>148</td>
<td>13.8</td>
<td>41</td>
<td>83</td>
<td>33.7</td>
</tr>
<tr>
<td>2007</td>
<td>1</td>
<td>9</td>
<td>188</td>
<td>14.6</td>
<td>41</td>
<td>84</td>
<td>37.5</td>
</tr>
<tr>
<td>2007</td>
<td>1</td>
<td>10</td>
<td>169</td>
<td>39.6</td>
<td>41</td>
<td>78</td>
<td>63.2</td>
</tr>
<tr>
<td>2007</td>
<td>1</td>
<td>11</td>
<td>160</td>
<td>19.2</td>
<td>37</td>
<td>66</td>
<td>86.9</td>
</tr>
<tr>
<td>2007</td>
<td>1</td>
<td>12</td>
<td>160</td>
<td>22.3</td>
<td>31</td>
<td>56</td>
<td>67.2</td>
</tr>
<tr>
<td>2007</td>
<td>1</td>
<td>13</td>
<td>166</td>
<td>11.7</td>
<td>27</td>
<td>55</td>
<td>40.4</td>
</tr>
</tbody>
</table>

- Real data set for LA area (South Coastal Air Quality Management District)
- 1,461 days of data (1/1/07- 12/31/10)
- Data described by Lopiano et al., obtained from them, [https://arxiv.org/abs/1502.03062](https://arxiv.org/abs/1502.03062)
- Original data sources: CARB for PM2.5 ([www.arb.ca.gov/aqmis2/aqdselect.php](http://www.arb.ca.gov/aqmis2/aqdselect.php)), CDPH for mortality counts, EPA for meteorological variables
- Download full data set from [http://cox-associates.com/CausalAnalytics/ LA_data_example.xlsx](http://cox-associates.com/CausalAnalytics/ LA_data_example.xlsx)
Alternative concepts of causality

- **Associational/attributive/(counterfactual)**
  - IARC: Regression, RR, burden-of-disease, PAR
  - Usually depends on untested assumptions
- **Predictive**: Causes help to predict their effects
  - Can be discovered and tested from data
  - Conditional independence tests, $X \rightarrow Y \rightarrow Z$
  - Granger tests, transfer entropy
- **Manipulative**: Changing causes changes effects
  - Randomized control trial (RCT)
  - Generalization/transportability
- **Mechanistic**: Changes propagate via networks of laws
  - Invariant laws (CPTs)
  - Composition of effects, well-behaved errors
Machine learning can help to avoid model-dependent conclusions and p-hacking

- Information-based algorithms: Automated, data-driven, minimal assumptions, empirically testable (usually)
  - Effects are *not conditionally independent* of their causes
  - Changes in causes *help to predict* changes in their effects
    - Granger causality for time series data; DAG models
  - Non-parametric methods minimize modeling assumptions
    - Trees
    - Bayesian networks
    - Causal directed acyclic graph (DAG) models
- Model ensembles address model uncertainty
  - RandomForest algorithm
  - Causal partial dependence plots

DAG = directed acyclic graph
Automated analysis with these methods is now practical: Enter data, click to analyze.
Automated analysis with these methods is now practical: Enter data, click to analyze
Automated analysis is now practical for all of the foregoing methods.

Executive Report:

What are the potential causal drivers of <AllCause75> in this data set?

The following were identified (by a Bayesian Network machine-learning algorithm) as potential causes of <AllCause75> in this data set:

Neighbors of <AllCause75> are: tmin, month, tmax

Potential causes of <AllCause75> are defined as its neighbors in a Bayesian Network.

The exposure variable [PM2.5] is NOT a significant predictor for [AllCause75] (p = 0.10) in a Quasi-Poisson regression model.

[tmin] is a significant predictor for [AllCause75] (p = 0.00) in a Quasi-Poisson regression model.

[month] is a significant predictor for [AllCause75] (p = 0.00) in a Quasi-Poisson regression model.

Significant predictors of <AllCause75> are defined here as those with regression coefficients significantly different from zero in a Quasi-Poisson regression model.

How important are these causal drivers?

From most to least important (using importance table), the relative importances of these potential causes are as follows:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Importance(%InC(MSE))</th>
</tr>
</thead>
<tbody>
<tr>
<td>month</td>
<td>168.28</td>
</tr>
<tr>
<td>tmin</td>
<td>62.27</td>
</tr>
<tr>
<td>tmax</td>
<td>34.07</td>
</tr>
<tr>
<td>PM2.5</td>
<td>5.83</td>
</tr>
</tbody>
</table>

A variable’s importance is measured here as the increase in mean squared error in predicting <AllCause75> if the variable is dropped.

How strongly does <PM2.5> predict or explain <AllCause75>?

Including <PM2.5> changes the percentage of explained variance in <AllCause75> from 40.25% to 40.80% in a randomForest analysis. Thus, including <PM2.5> as a predictor changes the

In multiple linear regression modeling, the percentage of explained variance (adjusted R-squared) in <AllCause75> is 31.44% when <PM2.5> is included and is 31.37% when <PM2.5> is dropped by about 0.07% in a multiple linear regression analysis.
Automated analysis is practical: trees and conditional probability tables
Automated analysis: Bayesian network

- Conditional probability table (CPT) or tree at each node
- Assumptions:
  - Causes are informative about their effects
  - Effects are not conditionally independent of causes
  - Direct causes are adjacent to their effects
  - Arrows reflect information, possible direct and indirect causal pathways
Automated non-parametric model: Bayesian network (BN)

Assumptions:
- Causes are informative about their effects.
  (Arrow directions here do not yet indicate causality)
- Effects are not conditionally independent of causes
- Direct causes are adjacent to their effects
- Arrows reflect information, possible direct and indirect causal pathways

Conditional probability table (CPT) or tree at each node
Exposure-response regression coefficient for PM2.5 as predictor of AllCause75 is significantly positive. Q: Why?

A: PM2.5 helps to correct model specification errors (errors in variables, month treated as a continuous predictor, omitted lagged daily temperatures)

Regression coefficients (and associations) mix direct, indirect, selection, confounder, and non-causal effects

Strong, consistent association ≠ evidence for predictive or manipulative causation
R packages and principles for identifying causal DAGs from data

• Conditional independence (constraint-based algorithms)
  – \textit{bnlearn, Tetrad, CompareCausalModels, dagitty} packages

• Likelihood principle (score-based algorithms)
  – Choose DAG model to maximize likelihood of data
  – Included among the algorithms in \textit{bnlearn} package

• Composition principle: If \( X \rightarrow Y \rightarrow Z \), then \( \frac{dz}{dx} = (\frac{dz}{dy})*(\frac{dy}{dx}) \)

• Granger/transfer entropy principle: Predictively useful information flows from causes to their effects over time
  – Transfer entropy, Yin & Yao, 2016, \url{www.nature.com/articles/srep29192}

• Model error specification principle
  – effect = f(cause) + error

• Homogeneity and invariance principles for causal CPTs
  – Li et al., 2015, \url{https://pdfs.semanticscholar.org/a051/9a2c6b85ca65d0df037142f550cf87d4e43f.pdf}
Automated analysis can be improved with causal knowledge, if available

Bayesian Network diagram.

Network discovered by bnlearn using model hc
An arrow between two variables shows that they are informative about each other.
Knowledge-based constraints
Potential p-hacking point, but controllable
Automated analysis can be improved with causal knowledge, if available

Interpretation:
• Now, arrows reflect information and knowledge-based constraints for predictive or manipulative causality
• CPT or tree at each node
• Causal CPTs or laws are invariant across studies
• Dynamic Bayesian networks (DBNs) include lagged values
Final models can be used to estimate causal input-output relations

Constrained automated non-parametric causal model ensembles can quantify causal relations
Automated estimation of causal input-output relation

Constrained automated non-parametric causal model ensemble result:
Mortality risk decreases slightly with same-day minimum temperature
Automated interpretation of statistical implications of a DAG model

Interpretation:
• Sound and complete inference algorithms generate all testable implications of DAG model learned from data
• Algorithms compute adjustment sets for estimating direct and total effects of changes in one variable on another for a given BN/DAG if its arrows and CPTs are causal

BN = Bayesian network
Summary: Machine learning helps avoid p-hacking and discover predictive causal relations

- Automated (but appropriate/intelligent) analyses can be carried out with current ML software for many real air pollution health effects data sets
  - Non-parametric
  - Information-based
  - Causal knowledge-constrained
  - Ensembles
- Enabled by existing R packages: randomForest, bnlearn, dagitty, CompareCausalNetworks, etc.
Some useful extensions

• Detecting omitted confounders
• Beyond DAGs
  – Allow for undirected arcs, cycles
• Transportability of results across settings
  – Appropriate generalization: Causal conditional probability tables (CPTs) are invariant, distributions of risk factors are not
• Combining results across studies
  – Different constraints from different studies
  – Causal CPTs are invariant across studies
Detecting hidden/omitted variables

Boston data
- Daily death counts in disjoint subpopulations are correlated
- Latent (hidden) variables affect both
- $F_{75} = \text{daily deaths among women 75 or older predicts}$
- $F_{75}$ predicts $M_{75}$
- $\text{All CVD 18to75}$ predicts (is informative about) both
Information-based causal discovery algorithms in perspective

• Philosophical underpinnings
  – Information flows from causes to effects over time
  – Tracking information flows enables data-driven causal discovery
  – Discovery = empirical constraints on possible models from observed information patterns in data
  • Differs from formulating a hypothesis and then testing it: Causal discovery imposes no \textit{a priori} hypotheses
  • Causal interpretation and orientation of arrows may require weak knowledge-based constraints
Practical aspects

• **Study design:** Ideally, track changes in exposures, covariates, and outcomes over time
  – *Data requirements* for causal discovery algorithms: Flexible (panel, time series, cross-sectional, etc.)

• **Assumptions:** Predictive causation + knowledge-based constraints provide a useful surrogate for manipulative causation

• **Model choices:** Learn tree ensembles, networks
  – Minimal assumptions, non-parametric, learned from data rather than assumed a priori
  – Use/compare multiple algorithms and principles

• **Sensitivity** to modeling choices: So far, causal model structure and estimates are robust to choice of algorithms
  – *CompareCausalNetworks* package
  – Model cross-validation
Caveats for information-based causal discovery algorithms

• Key assumptions:
  – *Data are available* to reveal information patterns and flows
    • Can be longitudinal or cross-sectional, many epidemiological and quasi-experimental (QE) designs suffice
  – *Effects are large enough to be detected* using non-parametric algorithms.
    • Power calculations reveal detection limits
    • (Causal Markov Condition, faithfulness, etc. useful but not essential)

• Limitations:
  – Unique identifiability from data not always possible → Must use multiple plausible models (model ensemble)
    • Arrow directions may be unclear, even in principle
    • Example: Income and air pollution
  – Predictive causation ≠ manipulative causation
  – Not yet well vetted for air pollution health effects research
    • Well vetted via Kaggle and other competitions in machine learning and causal learning communities
Conclusions

• Advice
  – Machine learning/information-based causal discovery is ready to apply to air pollution health effects data
    • Current software makes causal discovery relatively easy
  – Focus on predictive and manipulative causation (vs. other, e.g., associational/attributive or counterfactual, causation)
  – Focus on how well changes over time predict each other
    • Include at least 2 weeks of daily temperatures as lagged confounders in time series studies of daily mortality/morbidity
  – Use non-parametric model ensembles to avoid model specification errors, p-hacking, etc.

• Future research
  – Vet for air pollution health effects research
  – Compare information-based to potential outcomes methods in Kaggle-type competitions
Suggested readings

www.cox-associates.com/CausalAnalytics/

  – https://projecteuclid.org/euclid.ssu/1255440554


• Cox LA Jr., 2017. Do causal concentration-response functions exist? A critical review of associational and causal relations between fine particulate matter and mortality

• Cox LA Jr., 2017. Socioeconomic and air pollution correlates of adult asthma, heart attack, and stroke risks in the United States, 2010-2013.
Thanks!