Data-adaptive methods for causal inference with application to Mendelian randomization

Armitage Day, 7th November 2019
Data-adaptive methods and causal inference

• Machine learning is good at prediction
• Machine learning is good at pattern spotting (feature selection)
• Machine learning is not good at causal inference
• Use the strengths of machine learning as “pre-final estimation steps”
• Overlay robust causal thinking on top
Outline of talk

1. Introduction to Mendelian randomization
2. Variable selection for robustness to instrument invalidity
3. Variable selection for prioritizing causal determinants of disease
4. Clustering for identifying causal mechanisms
5. Investigating treatment effect heterogeneity
Observational data

- Correlation is not causation

- Observed associations between a risk factor and an outcome may result from:
  - Confounding
  - Reverse causation
Randomized trial

• In a randomized trial, participants are randomly assigned to a treatment group

• Random allocation ensures all potential confounders are equally distributed (on average) between the groups
  • Unmeasured confounding
  • Allocation of treatment gives a time-ordering
    • Reverse causation
An instrumental variable is:

1. Associated with the risk factor of interest
2. Not associated with any potentially confounding variable
3. Only associated with the outcome via the risk factor

Provides a natural experiment in observational data, similar to a randomized trial
Mendelian randomization

- Mendelian randomization is analogous to a randomized trial.
- An association between the genetic variant and the outcome is indicative of a causal effect of the risk factor on the outcome.
Mendelian randomization

- Genetic variants are particularly suitable candidate instrumental variables
  - Scientific knowledge of genetic function
  - Specific association with traits
  - Genetic sequence is determined at conception
- Mendelian randomization is the use of genetic variants as instrumental variables in observational data to obtain causal inferences
Causal estimation

- With a single instrument, the estimate of the causal effect of the risk factor on the outcome is:

\[
\hat{\beta}_{IV} = \frac{\hat{\beta}_{Y|Z}}{\hat{\beta}_{X|Z}}
\]

where \( \hat{\beta}_{Y|Z} \) is the coefficient from the regression of \( Y \) on \( Z \),
\( \hat{\beta}_{X|Z} \) is the coefficient from the regression of \( X \) on \( Z \),
\( \hat{\beta}_{IV} \) is the causal effect of \( X \) on \( Y \)
(the change in \( Y \) for a unit increase in \( X \))

- This relies on assumptions about linearity and lack of interactions
Motivation – failure in clinical trials

• Darapladib – inhibitor of lipoprotein-associated phospholipase A₂ (Lp-PLA₂)

• On the one hand, the drug worked – it did reduce the activity and concentration of Lp-PLA₂

• On the other hand, it failed to meet its primary endpoint in Phase 3 trial of 15,828 patients across median follow-up of 3.7 years ($$$)

• Could genetic knowledge have helped prevent this failed trial?
Genetic predictors of Lp-PLA2

Val279Phe (LoF)

<table>
<thead>
<tr>
<th>Trait</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lp PLA₂ activity</td>
<td>&lt;10⁻⁹⁵⁰</td>
</tr>
<tr>
<td>BMI</td>
<td>0.92</td>
</tr>
<tr>
<td>SBP</td>
<td>0.94</td>
</tr>
<tr>
<td>DBP</td>
<td>0.48</td>
</tr>
<tr>
<td>LDL-c</td>
<td>0.14</td>
</tr>
<tr>
<td>HDL-c</td>
<td>0.22</td>
</tr>
<tr>
<td>Triglycerides</td>
<td>0.35</td>
</tr>
<tr>
<td>C-reactive protein</td>
<td>0.43</td>
</tr>
<tr>
<td>Glucose</td>
<td>0.45</td>
</tr>
<tr>
<td>eGFR</td>
<td>0.31</td>
</tr>
</tbody>
</table>

Standardized mean difference per loss of function allele (95 % CI)

Val379Ala

<table>
<thead>
<tr>
<th>Trait</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lp PLA₂ activity</td>
<td>1.89x10⁻¹²</td>
</tr>
<tr>
<td>BMI</td>
<td>0.19</td>
</tr>
<tr>
<td>SBP</td>
<td>1.00</td>
</tr>
<tr>
<td>DBP</td>
<td>0.55</td>
</tr>
<tr>
<td>LDL-c</td>
<td>0.87</td>
</tr>
<tr>
<td>HDL-c</td>
<td>0.04</td>
</tr>
<tr>
<td>Triglycerides</td>
<td>0.85</td>
</tr>
<tr>
<td>C-reactive protein</td>
<td>0.74</td>
</tr>
<tr>
<td>Glucose</td>
<td>0.88</td>
</tr>
<tr>
<td>Insulin</td>
<td>0.65</td>
</tr>
<tr>
<td>eGFR</td>
<td>0.27</td>
</tr>
</tbody>
</table>

Standardized mean difference per allele (95 % CI)

Four LoF variants combined*

<table>
<thead>
<tr>
<th>Trait</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lp-PLA₂ activity</td>
<td>1.64x10⁻¹⁸</td>
</tr>
<tr>
<td>BMI</td>
<td>0.54</td>
</tr>
<tr>
<td>SBP</td>
<td>0.96</td>
</tr>
<tr>
<td>DBP</td>
<td>0.81</td>
</tr>
<tr>
<td>LDL-c</td>
<td>0.25</td>
</tr>
<tr>
<td>HDL-c</td>
<td>0.77</td>
</tr>
<tr>
<td>Triglycerides</td>
<td>0.99</td>
</tr>
<tr>
<td>C-reactive protein</td>
<td>0.38</td>
</tr>
<tr>
<td>Glucose</td>
<td>0.48</td>
</tr>
<tr>
<td>Insulin</td>
<td>0.85</td>
</tr>
<tr>
<td>eGFR</td>
<td>0.44</td>
</tr>
</tbody>
</table>

Standardized mean difference per loss of function allele (95 % CI)

Darapladib 160mg daily

<table>
<thead>
<tr>
<th>Trait</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lp-PLA₂ activity</td>
<td>&lt;10⁻⁹⁵⁰</td>
</tr>
<tr>
<td>SBP</td>
<td>0.03</td>
</tr>
<tr>
<td>LDL-c</td>
<td>0.84</td>
</tr>
<tr>
<td>HDL-c</td>
<td>0.04</td>
</tr>
<tr>
<td>Triglycerides</td>
<td>0.57</td>
</tr>
<tr>
<td>C-reactive protein</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Standardized mean difference per 160mg darapladib (95 % CI)
No association with disease risk
Genetic predictors of “bad” cholesterol
Polygenic Mendelian randomization

Ideal:

- Genetic association with LDL-cholesterol (mmol/L)

Reality:

- Association with CHD risk
- Association with HDL-cholesterol
Alternative causal pathways
Polygenic Mendelian randomization

Univariable Mendelian randomization (one risk factor)

• Regress SNP—outcome coefficients on SNP—risk factor coefficients, weighting for inverse-variance of SNP—outcome coefficients

\[ \hat{\beta}_{Yj} = \theta \hat{\beta}_{Xj} + \varepsilon_j, \quad \varepsilon_j \sim N(0, se(\hat{\beta}_{Yj})^2) \]

• Multivariable Mendelian randomization (several risk factors)

• Regress SNP—outcome coefficients on all SNP—risk factor coefficients, weighting for inverse-variance of SNP—outcome coefficients

\[ \hat{\beta}_{Yj} = \theta \hat{\beta}_{X1j} + \theta_{W1} \hat{\beta}_{W1j} + \theta_{W2} \hat{\beta}_{W2j} + \cdots + \varepsilon_j, \quad \varepsilon_j \sim N(0, se(\hat{\beta}_{Yj})^2) \]
## Multivariable Mendelian randomization

<table>
<thead>
<tr>
<th>Outcome</th>
<th>OR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fat mass index</strong></td>
<td></td>
</tr>
<tr>
<td>Aortic valve stenosis</td>
<td>1.46 (1.13-1.88)</td>
</tr>
<tr>
<td>Ischemic stroke</td>
<td>1.38 (1.19-1.60)</td>
</tr>
<tr>
<td>Transient ischemic attack</td>
<td>1.33 (1.15-1.54)</td>
</tr>
<tr>
<td>Atrial fibrillation</td>
<td>1.27 (1.14-1.40)</td>
</tr>
<tr>
<td>Abdominal aortic aneurysm</td>
<td>1.27 (0.93-1.71)</td>
</tr>
<tr>
<td>Thoracic aortic aneurysm</td>
<td>1.22 (0.69-2.16)</td>
</tr>
<tr>
<td>Heart failure</td>
<td>1.22 (1.06-1.41)</td>
</tr>
<tr>
<td>Peripheral artery disease</td>
<td>1.21 (1.05-1.39)</td>
</tr>
<tr>
<td>Subarachnoid hemorrhage</td>
<td>1.19 (0.99-1.44)</td>
</tr>
<tr>
<td>Deep vein thrombosis</td>
<td>1.18 (1.07-1.31)</td>
</tr>
<tr>
<td>Pulmonary embolism</td>
<td>1.14 (0.98-1.31)</td>
</tr>
<tr>
<td>Arterial hypertension</td>
<td>1.12 (1.04-1.20)</td>
</tr>
<tr>
<td>Intracerebral hemorrhage</td>
<td>1.09 (0.89-1.35)</td>
</tr>
<tr>
<td>Coronary artery disease</td>
<td>1.08 (1.00-1.17)</td>
</tr>
<tr>
<td><strong>Fat-free mass index</strong></td>
<td></td>
</tr>
<tr>
<td>Intracerebral hemorrhage</td>
<td>1.12 (0.85-1.48)</td>
</tr>
<tr>
<td>Arterial hypertension</td>
<td>1.08 (0.99-1.19)</td>
</tr>
<tr>
<td>Peripheral artery disease</td>
<td>1.05 (0.87-1.26)</td>
</tr>
<tr>
<td>Coronary artery disease</td>
<td>1.01 (0.91-1.12)</td>
</tr>
<tr>
<td>Deep vein thrombosis</td>
<td>0.98 (0.86-1.12)</td>
</tr>
<tr>
<td>Pulmonary embolism</td>
<td>0.95 (0.78-1.14)</td>
</tr>
<tr>
<td>Subarachnoid hemorrhage</td>
<td>0.92 (0.71-1.20)</td>
</tr>
<tr>
<td>Heart failure</td>
<td>0.92 (0.76-1.11)</td>
</tr>
<tr>
<td>Transient ischemic attack</td>
<td>0.88 (0.73-1.07)</td>
</tr>
<tr>
<td>Aortic valve stenosis</td>
<td>0.83 (0.59-1.16)</td>
</tr>
<tr>
<td>Ischemic stroke</td>
<td>0.75 (0.62-0.92)</td>
</tr>
<tr>
<td>Atrial fibrillation</td>
<td>0.75 (0.66-0.86)</td>
</tr>
<tr>
<td>Abdominal aortic aneurysm</td>
<td>0.64 (0.42-0.95)</td>
</tr>
<tr>
<td>Thoracic aortic aneurysm</td>
<td>0.63 (0.30-1.32)</td>
</tr>
</tbody>
</table>

OR (95% CI) per 1 kg/m² increase in fat mass and fat-free mass indices
Multivariable MR with lasso

Univariable:  \[ \hat{\beta}_{Yj} = \theta \hat{\beta}_{Xj} + \varepsilon_j, \quad \varepsilon_j \sim N(0, se(\hat{\beta}_{Yj})^2) \]

\[ \hat{\theta}_{IVW} = \arg \min_\theta \sum_j se(\hat{\beta}_{Yj})^{-2} (\hat{\beta}_{Yj} - \theta \hat{\beta}_{Xj})^2 \]

Multivariable:

\[ \hat{\beta}_{Yj} = \theta \hat{\beta}_{X1j} + \theta W \hat{\beta}_{Wj} + \varepsilon_j, \quad \varepsilon_j \sim N(0, se(\hat{\beta}_{Yj})^2) \]

\[ \hat{\theta} = \arg \min_\theta \sum_j se(\hat{\beta}_{Yj})^{-2} (\hat{\beta}_{Yj} - \theta \hat{\beta}_{Xj} - \theta W \hat{\beta}_{Wj})^2 \]

where \( j \) indexes genetic variants and \( k \) indexes risk factors
Variable selection for robustness

- If the number of alternative risk factors $<$ number of variants
  - Ignoring alternative risk factors leads to bias
  - Including all risk factors in the model can lead to inefficiency
  - Variable selection approach is optimal balance
- If the number of alternative risk factors $>$ number of variants
  - Must use regularization
  - Lasso
  - Double selection
  - Inference post-selection
Variable selection for prioritizing causal risk factors

• Lipids cause coronary artery disease

• But which are the causal risk factors?

• Take genetic associations for 148 variants with 48 lipid measurements

• Perform variable selection

\[ \hat{\beta}_{Yj} = \theta_1 \hat{\beta}_{M_{1j}} + \theta_2 \hat{\beta}_{M_{2j}} + \theta_3 \hat{\beta}_{M_{3j}} + \cdots + \varepsilon_j \]

148 genetic variants associated with lipids from the Global Lipid Genetics Consortium

NMR metabolites as potential risk factors which are highly correlated
Variable selection for prioritizing causal risk factors

• Bayesian variable selection

• Consider all single risk factor models, all pairs, all triples, etc
  • For many risk factors, use stochastic search
  • For each set of risk factors, calculate a Bayes factor

• Calculate posterior model probability, marginal inclusion probability of each risk factor

Apolipoprotein B (ApoB)
Clustering heterogeneous variants

Association with CHD risk

Association with HDL-cho
Likelihood-based clustering approach

- Suppose there are $K$ (substantive) clusters in the data
- The causal estimate for variant $j$ in cluster $k$ is normally distributed about a cluster mean $\theta_k$ with known variance
- Can construct a likelihood as a mixture of normal distributions
- We also include a null cluster and a junk cluster
- Null cluster: cluster mean is zero
- Junk cluster: causal estimates are drawn from a heavy-tailed distribution (generalized t-distribution)
Likelihood

\[
p(\hat{\theta}_j \mid \Theta, \hat{\sigma}^2_j) = \sum_{k=0}^{K+1} p(\hat{\theta}_j, z_j = k \mid \Theta, \hat{\sigma}_i^2) \\
= \sum_{k=0}^{K+1} p(z_j = k)p(\hat{\theta}_j \mid \Theta, \hat{\sigma}^2_j, z_j = k) \\
= \pi_0 \phi(\hat{\theta}_j \mid 0, \hat{\sigma}_j^2) + \sum_{k=1}^{K} \pi_k \phi(\hat{\theta}_j \mid \theta_k, \hat{\sigma}_j^2) + \pi_{K+1} T(\hat{\theta}_j)
\]

- Maximize using expectation-maximization algorithm
- Use BIC to determine number of clusters
Treatment effect heterogeneity

• Instrumental variable estimate represents a population-averaged causal effect

• What if the true effect varies for different individuals?

• Take a genetic score that is a proxy for LDL-cholesterol lowering (variants in HMGCR gene region)

• Find genetic variants that have statistical interaction in the association between LDL-cholesterol and pharmacomimetic score
Treatment effect heterogeneity

• Assess gene—gene interactions for each variant in turn
• Take all variants that have interactions at $p<10^{-5}$ (say)
• Construct a tree using two-thirds of the data (training dataset)
• Estimate the causal effects in the validation dataset
• Calculate the variance in the node-specific causal estimates calculated in the validation dataset using the tree from the training dataset
One tree
Treatment effect heterogeneity - results

- Training Set SD
- Validation Set SD
- Permutated Validation Set SD
Treatment effect heterogeneity - results

![Graphs showing treatment effect heterogeneity results with interaction thresholds and node sizes.](image)
Treatment effect heterogeneity - results
Treatment effect heterogeneity - results

• In this case, no more heterogeneity in treatment effects than would be expected by chance alone

• If there was heterogeneity, could identify groups who respond more/less strongly to treatment

• Create a “polygenic treatment score” – a moderating variable that predicts response to treatment
  • Identify those in the 98th percentile for response to statins
  • Allow better treatment decisions (stratified/personalized medicine)
1. Introduction to Mendelian randomization

2. Variable selection for robustness to instrument invalidity

3. Variable selection for prioritizing causal determinants of disease

4. Clustering for identifying causal mechanisms

5. Investigating treatment effect heterogeneity
Conclusions

• Causal assumptions are key – “no lipstick on a pig”
• Use data-adaptive methods for the non-causal aspect of the model
• Need to question our causal assumptions – we do this!
  • Robustness to outliers/influential points
• Interested to discuss further!
References

Acknowledgements

- MRC Biostatistics Unit
  - Andrew Grant
  - Chris Foley
  - Apostolos Gkatzionis
  - Verena Zuber (now Imperial)
  - Zhi Ming Xu

- Cardiovascular Epidemiology Unit
  - Amy Mason
  - Jessica Rees
  - plus many collaborators