Causal Discovery from Pediatric Infectious Disease Protein Biomarker Data

Subramani Mani, MBBS (MD), PhD

Conflict of Interest

- None

Roadmap

- Introduction to causal discovery from observational data
- Y structure theorem
- BLCD algorithm
- Results
  - Late onset neonatal sepsis biomarkers
  - Pediatric infections biomarkers
- Discussion
  - Effect of hidden variables
  - Biomarker causal modeling challenges
- Future directions

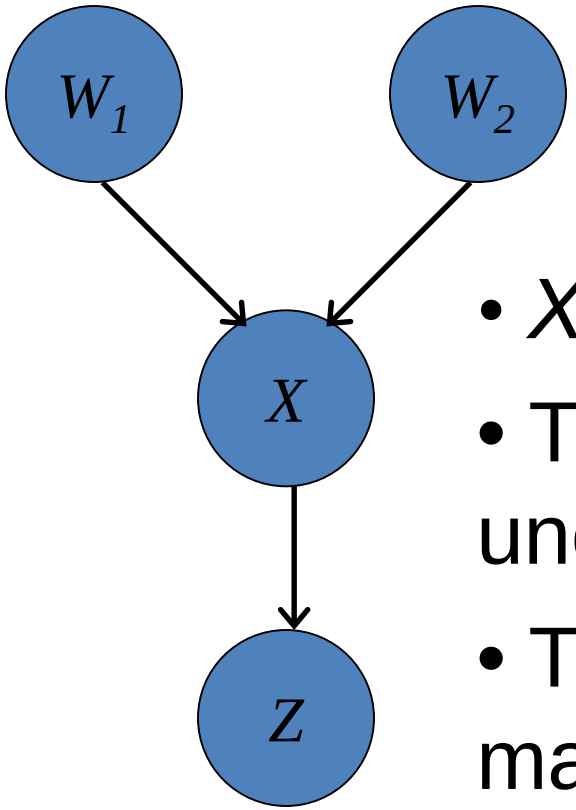
Why Causality?

- Causal knowledge has the potential to tell us the effects of manipulation of the world.
- Gives us the insight to plan interventions leading to desirable outcomes.

Learning Causality from Observational Data

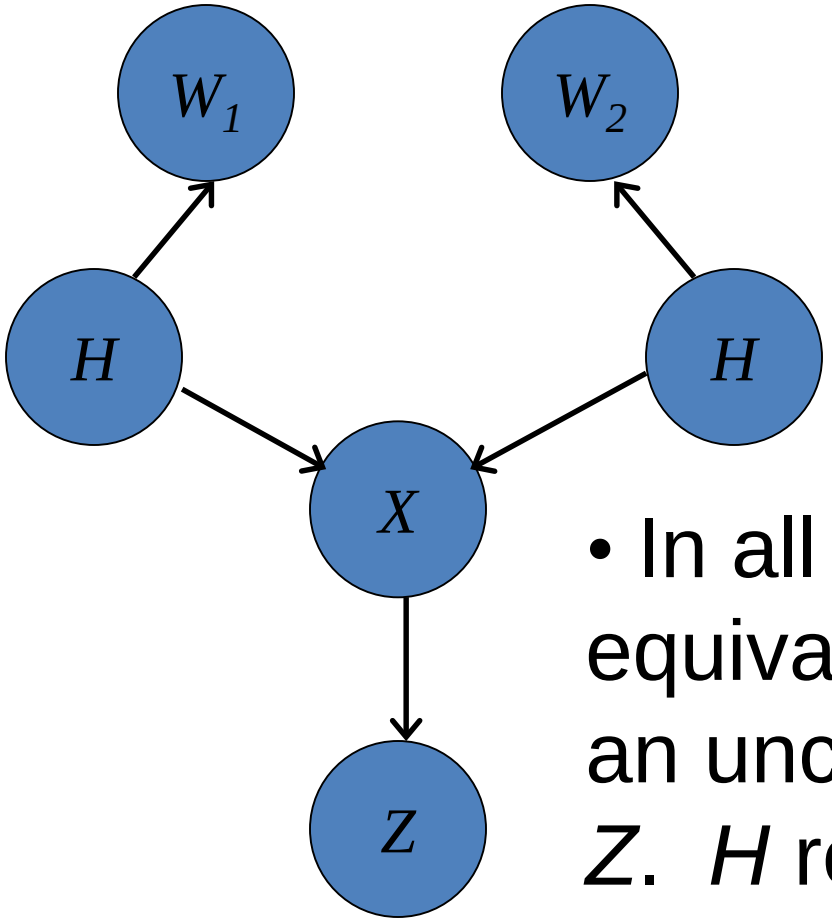
- Challenging due to the presence of hidden confounders.
- Use of a causal Bayesian network framework
- Two basic search approaches are typically employed
  - Constrained based
  - Score based

Y Structure



- *X* causally influences *Z*.
- The arc from *X* to *Z* is unconfounded.
- The effect on *Z* of manipulating *X* is just P(*Z* | *X*)

Y Equivalent Structure



- In all DAGs independence equivalent to Y structure, *X* is an unconfounded ancestor of *Z*. *H* represents a hidden variable.

Y Structure Sufficiency Theorem

- The Y structure sufficiency theorem indicates that local Bayesian causal discovery using Y structures is possible (under assumptions), even when the data generating process is assumed to be a causal Bayesian network with hidden variables. See [1] for a proof of correctness.

[1]. Mani, S., Spirtes, P., & Cooper, G.F. (2006). A theoretical study of Y structures for causal discovery. UAI 2006, pp. 314-323. Corvallis, OR: AUAI Press.

Markov Blanket (MB)

- The MB of a node *X* in a causal Bayesian network G is the union of
  - The set of parents of *X*
  - The children of *X*
  - The parents of the children of *X*
- Conditioning on the MB of a node *X*, makes *X* independent of all the other nodes in G

BLCD—A Bayesian Local Causal Discovery Algorithm

- Assumptions for Causal Discovery
  - Markov assumption.
  - Faithfulness assumption.
  - There is a data generating Bayesian network B that has a structure S that is Markov and faithful to the distribution that B represents.

BLCD Scoring Measure

- Total number of DAGs G_i on four measured variables is 543.
- Estimates the unconfounded causal influence of X on Z, where D denotes observational data.

$$P(X \Rightarrow Z \mid \mathrm{D}) \cong \frac{\mathrm{Score}(\mathrm{Y\ structure} \mid \mathrm{D})}{\sum_{i=1}^{543} \mathrm{Score}(\mathrm{G}_i \mid \mathrm{D})} \qquad (1)$$

where $\Rightarrow$ denotes unconfounded causation

BLCD: Steps

1. Derive the MB of each node $X \in \mathbf{V}$. Let **B** denote the MB of X.
   - A greedy forward and backward heuristic search is used for this step.
2. Update **B**. If X is in the MB of Z, add Z to the MB of X.
3. Pick $W1$, $W2$, Z from **B**. Add X to get a set of 4 variables.
4. Derive $P(X \rightarrow Z \mid D)$.
5. Generate output. If $P(X \rightarrow Z \mid D) > t$, where t is a user-set threshold, then output $X \rightarrow Z$ as plausibly causal.

BLCD Algorithm Summary

- Input
  - An observational dataset *D* over a set of observed random variables **V**.
- Output
  - Probabilities of the form $X \rightarrow Z$ for which *X* is an unconfounded cause of *Z* (under assumptions). See [1] for a proof of correctness.

[1]. Mani, S., Spirtes, P., & Cooper, G.F. (2006). A theoretical study of Y structures for causal discovery. UAI 2006, pp. 314-323. Corvallis, OR: AUAI Press.

Real-World datasets

- Hypothesis: We can discover meaningful cause and effect relationships from biomedical observational data allowing for unmeasured or hidden variables
- Neonatal sepsis biomarker dataset
- Pediatric infections biomarker dataset
- To identify cause-effect relationships unconfounded by unmeasured variables using the BLCD causal discovery algorithm

Neonatal Sepsis Dataset

- Study sample
  - 127 infants enrolled over a five year period (2007-2012)
  - 39 cases and 88 controls
- Inclusion criteria
  - Gestational age of ≤32 weeks
  - Birth weight ≤1500 grams
  - Postnatal age of ≥5 days

Proteomic Data

Patient	Protein-1	Protein-2	Protein-3	.	Protein-90	Diagnosis
1	.	.	.	.	.	Sepsis
2	.	.	.	.	.	Control
3	.	.	.	.	.	Control
4	.	.	.	.	.	Sepsis
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
127	.	.	.	.	.	Control

Sample Characteristics (n=127)

Variable Description	Statistics
Birth weight (grams): median (Q1,Q3)	895 (718,1145)
Gestational age (weeks): median (Q1,Q3)	27 (25,29)
Maternal age: median (Q1,Q3)	26 (21,32)
Male: n (%)	62 (49%)
Sepsis negative: n (%)	88 (69%)
Race (White): n (%) Black Others Unknown	88 (69%) 21 (16.5%) 17 (13.5%) 1 (1%)

Neonatal Sepsis Data: Output of BLCD

Cause	Effect	Plausible
1. A2Macro	CK-MB	?
2. IL-15	A2Macro	?
3. IL-15	Thrombopoietin	?
4. IL-15	IL-5	?
5. CK-MB	Myoglobin	?
6. CD40-L	TBG	?
7. A2Macro	Fibrinogen	?
8. TNFR2	CRP	Yes
9. IL-15	IL-12p40	Yes

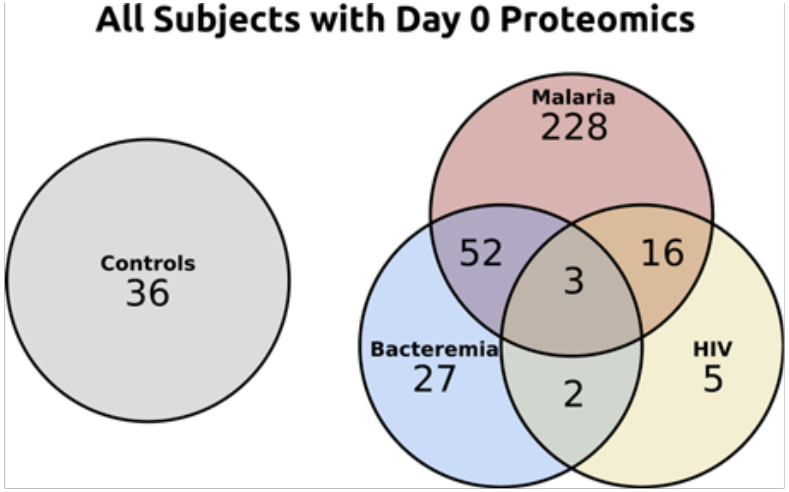
Neonatal Sepsis Data: Evaluation Based on Published Literature

- TNFR2 causally influencing CRP is supported by literature with TNFR2 becoming upregulated prior to CRP [2]
- IL-15 causally influencing IL-12p40 is also supported by literature [3]

[2] Doellner H, Arntzen KJ, Haereid PE, Aag S, Brubakk A-M, Austgulen R. Increased serum concentrations of soluble tumor necrosis factor receptors p55 and p75 in early onset neonatal sepsis. Early human development 1998;52(3):251-61
[3] Parayath KE, Harrison TS, Levitz SM. Effect of interleukin (IL)-15 priming on IL-12 and interferon-γ production by pathogen-stimulated peripheral blood mononuclear cells from human immunodeficiency virus-seropositive and-seronegative donors. Journal of Infectious Diseases 2000;181(2):733-36

Pediatric Infection Biomarkers Dataset

- Study sample
  - 1655 children were enrolled over a period of ten years from Kenya
  - First visit before the age of 14 months
  - Follow up post enrollment for 36 months
  - Protein biomarkers using 25-plex human inflammatory array from day 0 available for 369 children



Disease conditions included in the 25-plex human inflammatory array.

Pediatric Infections Data: Output of BLCD

Cause	Effect	Plausible
1. IL-4	L-2R	Yes
2. IL-4	Eotaxin	Yes
3. IL-4	MCP-1	Yes
4. IFN-a	IP-10	Yes
5. MIP-1B	IL-1Ra	?
6. MIP-1B	IL-5	?
7. MIP-1B	IL-7	?

Pediatric Infections Data: Evaluation Based on Published Literature

- The first four relationships are biologically meaningful [4]
- The “causal” relationship between MIP-1B and IL-1Ra, IL-5, IL-7 need further investigation

[4] Perkins DJ, Were T, Davenport GC, Kempaiah P, Hittner JB, Ong'echa JM. Severe malarial anemia: innate immunity and pathogenesis. International journal of biological sciences 2011;7(9):1427

Limitations

- The method assumes a directed acyclic graph structure as causal models.
- We can discover only causal relationships represented in nature as Y structures.
- Evaluation done on only two biomedical datasets.
- Hidden variables modeled implicitly.

Future Work

- Using causal discovery methods to generate clinical practice guidelines from data
- Discovering causal influences from large biomedical datasets
- Mechanistic modeling of other disease linked biomarkers
- Causal discovery from temporal data

Acknowledgements

Causal Modeling Methods

- Collaborators
  - Greg Cooper
  - Peter Spirtes
  - Alex Statnikov
  - Yukun Chen

Causal Modeling Studies

- Collaborators
  - Robin Ohls
  - DJ Perkins
  - Cristian Bologa
  - Daniel Cannon

Funding

- Thrasher foundation (PI Ohls)
- CTSC Pilot Neonatal Sepsis (PI Mani)
- CTSC Pilot Pediatric Infections (PI Mani)

Thanks!

Questions?? Comments!!

Learning CBNs: Bayesian Scoring of Models

- We can derive the posterior probability of a BN structure B_S as follows:

$$P(B_S \mid D) = \frac{P(B_S, D)}{P(D)} = \frac{P(B_S, D)}{\sum_{B_S} P(B_S, D)}$$

$$P(B_S, D) = P(B_S) \int P(D \mid B_S, \theta_{B_S}) P(\theta_{B_S} \mid B_S) d\theta_{B_S}$$

Learning CBNs: BDe Scoring Measure

$$P(B_S, D) = P(B_S) \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})}$$

where:

$P(B_S)$ is the prior probability of the CBN B_S.

n is the number of nodes in the CBN.

Γ is the gamma function.

q_i is the number of unique instantiations of the parents of node i in database D. If node i has no parents, then $q_i = 1$.

r_i denotes the number of discrete values of node i.

N_{ijk} is the number of instances in D that node i has value k and the parents of i have the instantiation denoted by j.

N_{ij} is the number of instances in D that the parents of node i have the instantiation denoted by j.

α_{ij} and α_{ijk} represent prior sample sizes.

Neonatal Sepsis Potential Protein Biomarkers

1. Adiponectin	31. Glutathione S-Transferase	61. Lymphotactin
2. Alpha-1 Antitrypsin	32. GM-CSF	62. MCP-1
3. Alpha-2 Macroglobulin	33. Growth Hormone	63. MDC
4. Alpha-Fetoprotein	34. Haptoglobin	64. MIP-1alpha
5. Apolipoprotein A1	35. ICAM-1	65. MIP-1beta
6. Apolipoprotein CIII	36. IFN-gamma	66. MMP-2
7. Apolipoprotein H	37. IgA	67. MMP-3
8. Beta-2 Microglobulin	38. IgE	68. MMP-9
9. BDNF	39. IGF-1	69. Myeloperoxidase
10. C Reactive Protein	40. IgM	70. Myoglobin
11. Calcitonin	41. IL-10	71. PAI-1
12. Cancer Antigen 125	42. IL-12p40	72. PAPP-A
13. Cancer Antigen 19-9	43. IL-12p70	73. PSA-Free
14. Carcinoembryonic Antigen	44. IL-13	74. Prostatic Acid Phosphatase
15. CD40	45. IL-15	75. RANTES
16. CD40 Ligand	46. IL-16	76. Serum Amyloid P
17. Complement 3	47. IL-18	77. SGOT
18. Creatine Kinase-MB	48. IL-1alpha	78. SHBG
19. EGF	49. IL-1beta	79. Stem Cell Factor
20. EN-RAGE	50. IL-1ra	80. Thrombopoietin
21. ENA-78	51. IL-2	81. TSH
22. Endothelin-1	52. IL-3	82. Thyroxine Binding Globulin
23. Eotaxin	53. IL-4	83. TIMP-1
24. Erythropoietin	54. IL-5	84. Tissue Factor
25. Factor VII	55. IL-6	85. TNF RII
26. Fatty Acid Binding Protein	56. IL-7	86. TNF-alpha
27. Ferritin	57. IL-8	87. TNF-beta
28. FGF basic	58. Insulin	88. VCAM-1
29. Fibrinogen	59. Leptin	89. VEGF
30. G-CSF	60. Lipoprotein (a)	90. von Willebrand Factor

Pediatric Infections Potential Protein Biomarkers

Abbreviation	Full Name
IL-1B	Interleukin-1 beta
IL-1Ra	Interleukin-1 receptor antagonist
IL-2	Interleukin-2
IL-2R	Interleukin-2 receptor
IL-4	Interleukin-4
IL-5	Interleukin-5
IL-6	Interleukin-6
IL-7	Interleukin-7
IL-8	Interleukin-8
IL-10	Interleukin-10
IL-12p40/70	Interleukin-12 Subunit p40 divided by Interleukin-12 Subunit p70
IL-13	Interleukin-13
IL-15	Interleukin-15
IFN-a	Interferon alpha
IL-17	Interleukin-17
TNF-a	Tumor necrosis factor alpha
IFN-y	Interferon gamma
GM-CSF	Granulocyte-macrophage colony stimulating factor
IP-10	Interferon gamma-induced protein 10
MIP-1a	Macrophage inflammatory protein 1 alpha
MIP-1B	Macrophage inflammatory protein 1 beta
MIG	Macrophage induced by gamma interferon
Eotaxin	Eotaxin
RANTES	Regulated on activation, normal T cell expressed and secreted
MCP-1	Monocyte chemotactic protein 1

Note: Eotaxin is CCL11 (eotaxin-1) and the IL-12p40/p70 measures p40, the p40-p40 (homodimer), and IL-12 p70 (p35 and p40 subunits).