Learning interpretable causal networks from observational data

Marcel DA CÂMARA RIBEIRO-DANTAS, PhD

mribeirodantas@seqera.io

Causal τ working group seminar
January 12th, 2023

Commission européenne

MARIE CURIE ACTIONS

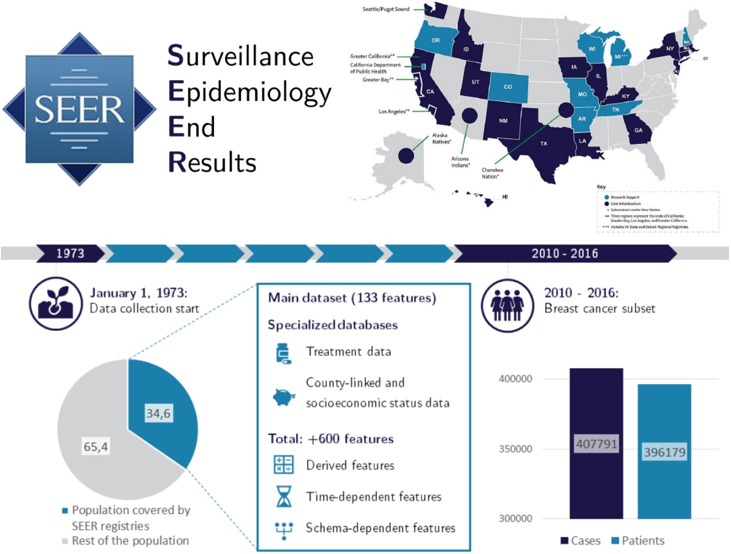
institut Curie

SORBONNE UNIVERSITÉ

Outline

SEER

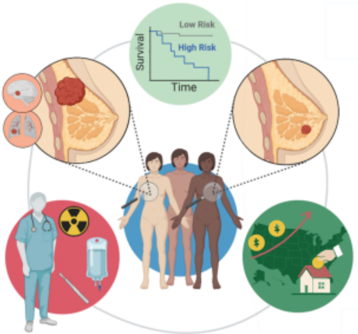
The Surveillance, Epidemiology, and End Results (SEER) Program



SEER

Breast Cancer (BC)

- Cancer originated in breast cells
- Types and sub-types depend on cell characteristics
- Most common invasive cancer in women
- Most common cancer-related cause of death in women
- Increasing prevalence since the 70's
- Specific variables for BC in SEER



5-year relative survival rates for breast cancer

These numbers are based on women diagnosed with breast cancer between 2011 and 2017.

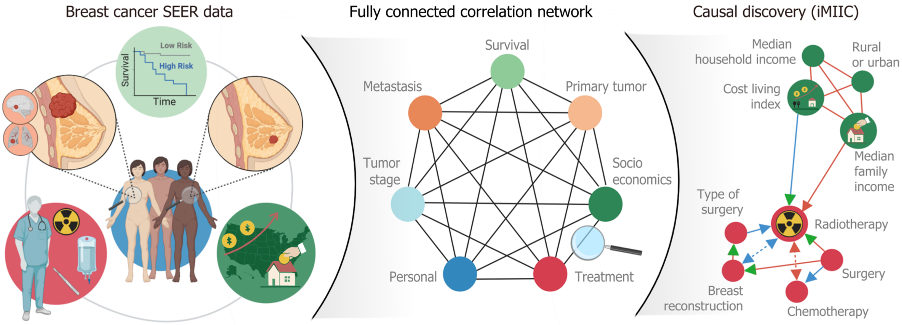
SEER Stage	5-year Relative Survival Rate
Localized*	99%
Regional	86%
Distant	29%
All SEER stages combined	90%

*Localized stage only includes invasive cancer. It does not include ductal carcinoma in situ (DCIS).

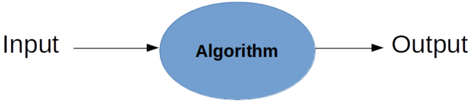
American Cancer Society, 2022.

Causal Discovery and iMIIC

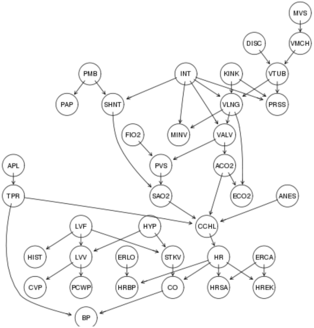
Network inference

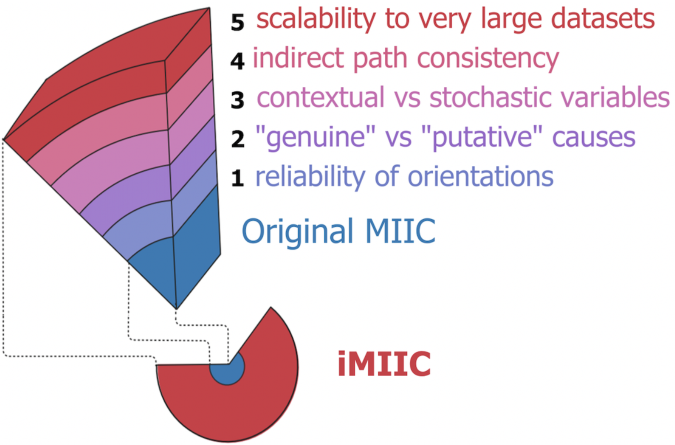


Disentangling direct from indirect relations between variables, including *or* excluding cause-effect relationships.



	CVP	PCWP	HIST	TPR
1	0.175098608713597	0.288900742307305	0.55428269	0.857808550121263
2	0.0933099433314055	0.218888618983328	0.47473369	0.332442865008488
3	0.690925023518503	0.861214176984504	0.25389608	0.849817770067602
4	0.572590821655467	0.549840433290228	0.15323819	0.715732422890142
5	0.857235474744812	0.255593683104962	0.49391256	0.37724070623517
6	0.590208335081115	0.367558936588466	0.62587462	0.933418722823262
7	0.816242689266801	0.526696094544604	0.47205955	0.651990963146091
8	0.76507281861268	0.835657971445471	0.96377760	0.984965795883909
9	0.885613681515679	0.196845271624625	0.50106454	0.293295677984133
10	0.941809490323067	0.956555964192376	0.07710378	0.941999231465161
11	0.685077040921897	0.517504557501525	0.49092264	0.731579512590542
12	0.0605227875057608	0.759360220748931	0.69840481	0.663990918546915
13	0.19431169051677	0.477279279148206	0.67160601	0.996502364054322
14	0.614625208079815	0.360529601573944	0.02014737	0.375805815681815
15	0.700897089438513	0.777111812029034	0.56314651	0.849968496710062





5 scalability to very large datasets
4 indirect path consistency
3 contextual vs stochastic variables
2 "genuine" vs "putative" causes
1 reliability of orientations
Original MIIC
iMIIC

Search & Score (Scoring function ϕ)

Find the graph $\mathcal{G}$ that **maximizes** the score $\phi_{\mathcal{G}}$ (.e.g, Likelihood)

Super-exponential space of networks, only $\rightarrow$ (i.e. assumes causality)

Causal Discovery and iMIIC

Search & Score and constraint-based methods

Search & Score (Scoring function ϕ)

Find the graph $\mathcal{G}$ that **maximizes** the score $\phi_{\mathcal{G}}$ (.e.g, Likelihood)

Super-exponential space of networks, only $\rightarrow$ (i.e. assumes causality)

Constraint-based (Conditional independences)

Broader network class including $—$ $\rightarrow$ $\leftrightarrow$ Signature of causality: $X \rightarrow Z \leftarrow Y$

Interpretability and **sampling noise issues** (Spurious conditional independence)

Causal Discovery and iMIIC

Search & Score and constraint-based methods

Search & Score (Scoring function ϕ)

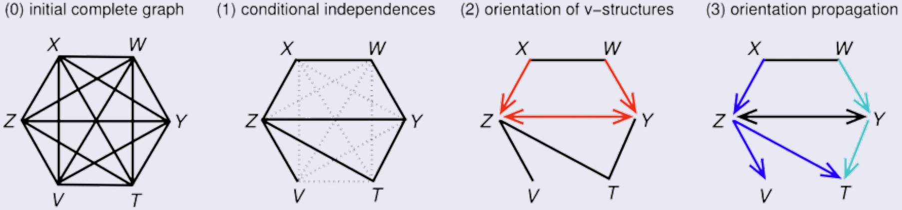
Find the graph $\mathcal{G}$ that **maximizes** the score $\phi_{\mathcal{G}}$ (.e.g, Likelihood)

Super-exponential space of networks, only $\rightarrow$ (i.e. assumes causality)

Constraint-based (Conditional independences)

Broader network class including — $\rightarrow$ $\leftrightarrow$ Signature of causality: $X \rightarrow Z \leftarrow Y$

Interpretability and **sampling noise issues** (Spurious conditional independence)

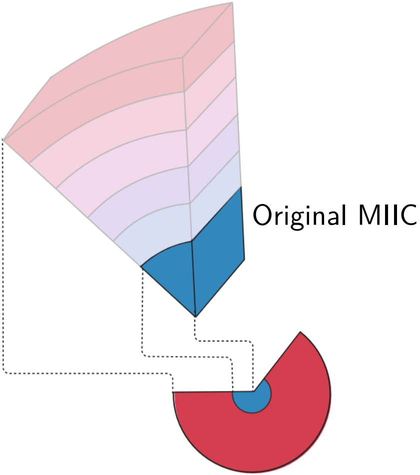


without latent variables: PC (Spirtes 1991), IC (Pearl 1991)

with latent variables: FCI (Spirtes 1999), AFCI (Spirtes 2001), RFCI (Colombo 2012)

Causal Discovery and iMIIC

Original MIIC



Original MIIC based on the **3off2** scheme

Robust constraint-based approaches to **finite dataset** (N), based on **iterative collection** of information **contributors** $\{a_i\}_n$ to $I(x;y)$

$$I(x;y|\{a_i\}_n) = I(x;y) - I(x;y;a_1) - I(x;y;a_2|a_1) - \cdots - I(x;y;a_n|\{a_i\}_{n-1})$$

Original MIIC based on the **3off2** scheme

Robust constraint-based approaches to **finite dataset** (N), based on **iterative collection** of information **contributors** $\{a_i\}_n$ to $I(x;y)$

$$I(x;y|\{a_i\}_n) = I(x;y) - I(x;y;a_1) - I(x;y;a_2|a_1) - \cdots - I(x;y;a_n|\{a_i\}_{n-1})$$

Conditional independence (Including finite size correction)

$$I'(x;y|\{a_i\}_n) = I(x;y|\{a_i\}_n) - \frac{1}{2} k_{x;y|\{a_i\}_n} \frac{\log N}{N} \leqslant 0$$

Original MIIC based on the **3off2** scheme

Robust constraint-based approaches to **finite dataset** (N), based on **iterative collection** of information **contributors** $\{a_i\}_n$ to $I(x;y)$

$$I(x;y|\{a_i\}_n) \;=\; I(x;y) - I(x;y;a_1) - I(x;y;a_2|a_1) - \cdots - I(x;y;a_n|\{a_i\}_{n-1})$$

Conditional independence (Including finite size correction)

$$I'(x;y|\{a_i\}_n) \;=\; I(x;y|\{a_i\}_n) - \frac{1}{2} k_{x;y|\{a_i\}_n} \frac{\log N}{N} \leqslant 0$$

3-point Multivariate Information (Positive or Negative)

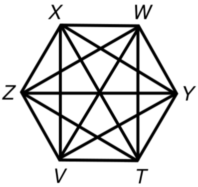
$$I'(x;y;z|\{a_i\}) = I'(x;y|\{a_i\}) - I'(x;y|\{a_i\},z)$$

(1) Affeldt, Isambert; UAI 2015. (2) Affeldt, Verny, Isambert; BMC Bioinformatics, 2016. (3) Verny, Sella, Affeldt, Singh, Isambert; PLOS Comput Biology, 2017. (4) Sella, Verny, Uguzzoni, Affeldt, Isambert; Bioinformatics, 2018. (5) Cabeli, Verny, Sella, Uguzzoni, Verny, Isambert; PLOS Comput Biology 2020

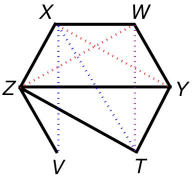
Causal Discovery and iMIIC

Original MIIC algorithm

(0) Complete graph



(1) Remove edges $I'(X_1; X_2|\{A_i\}) \simeq 0$



$I'(X, T|Y, Z) \simeq 0$

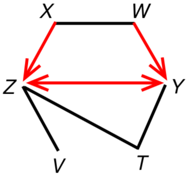
$I'(X; Y|W) \simeq 0$

$I'(X; V|Z) \simeq 0$

$I'(W; Z|X) \simeq 0$

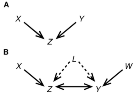
$I'(W; T|Y, Z) \simeq 0$

(2) Orient V-structures (P_{head})

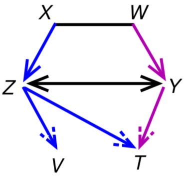


$I'(X; Y; Z|W) < 0$

$I'(W; Z; Y|X) < 0$



(3) Non-v-structures (P_{tail})



$I'(X; V; Z) > 0$

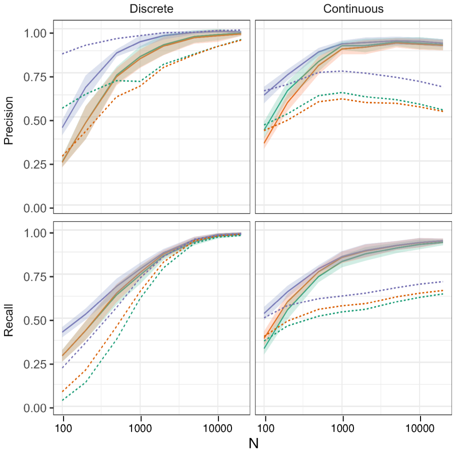
$I'(X; T; Z|Y) > 0$

$I'(W; T; Y|Z) > 0$

$$P_h(X \to \underline{Z}) = P_h(\underline{Z} \leftarrow Y) = \frac{1+e^{N\,I'(X;Y;Z|\{A_i\})}}{1+3e^{N\,I'(X;Y;Z|\{A_i\})}} \quad P_t(\underline{Z} \to V) = \frac{1}{1+e^{-N\,I'(X;Z;V|\{A_i\})}} P_h(X \to \underline{Z})$$

Causal Discovery and iMIIC

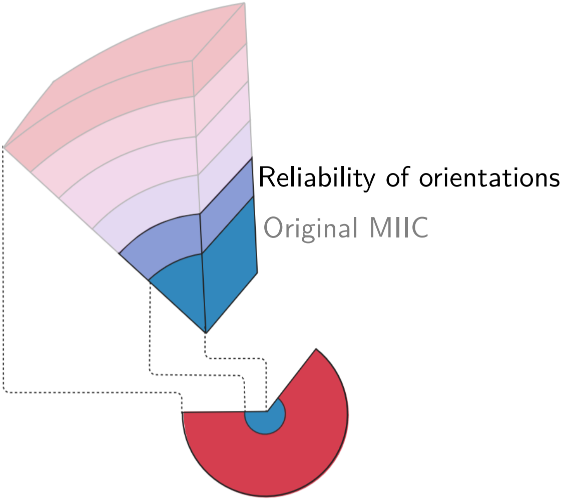
Original MIIC algorithm



Original MIIC —— *vs* PC · · ·   Skeleton   Oriented-edge-only   CPDAG

Causal Discovery and iMIIC

Reliability of Orientations



Consistent *versus* inconsistent V-structures

- If $I'(x;y|\{a_i\}) < 0,\ I'(x;y|\{a_i\},z) > 0$
  $\implies I'(x;y;z|\{a_i\}) = I'(x;y|\{a_i\}) - I'(x;y|\{a_i\},z) < 0$
  $\implies x \to z \leftarrow y$ (Consistent)
- If $I'(x;y|\{a_i\}) < I'(x;y|\{a_i\},z) < 0$
  $\implies I'(x;y;z|\{a_i\}) = I'(x;y|\{a_i\}) - I'(x;y|\{a_i\},z) < 0$
  $\implies x \to z \leftarrow y$ (Inconsistent)

Consistent *versus* inconsistent V-structures

- If $I'(x;y|\{a_i\}) < 0,\ I'(x;y|\{a_i\},z) > 0$
  $\implies I'(x;y;z|\{a_i\}) = I'(x;y|\{a_i\}) - I'(x;y|\{a_i\},z) < 0$
  $\implies x \rightarrow z \leftarrow y$ (Consistent)
- If $I'(x;y|\{a_i\}) < I'(x;y|\{a_i\},z) < 0$
  $\implies I'(x;y;z|\{a_i\}) = I'(x;y|\{a_i\}) - I'(x;y|\{a_i\},z) < 0$
  $\implies x \rightarrow z \leftarrow y$ (Inconsistent)

More conservative orientations by rectifying negative MI* and CMI

Before rectification $\quad I'(x;y|\{a_i\}) < I'(x;y|\{a_i\},z) < 0.$

After rectification $\quad I'(x;y|\{a_i\}) = I'(x;y|\{a_i\},z) = 0.$

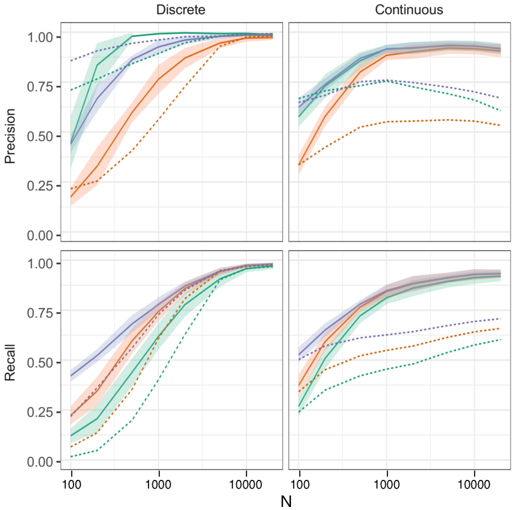
$\implies I'(x;y;z|\{a_i\}) = I'(x;y|\{a_i\}) - I'(x;y|\{a_i\},z) = 0$

$\implies x - z - y$ remains non-oriented.

* as $I'(X;Y) = \sup_{\mathcal{P},\mathcal{Q}} I'([X]_{\mathcal{P}};[Y]_{\mathcal{Q}}) \geqslant I'([X]_1;[Y]_1) = 0$

Causal Discovery and iMIIC

Orientation in iMIIC

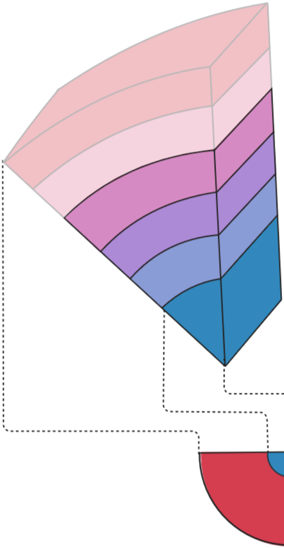


Conservative MIIC — *vs* Conservative PC · · · Skeleton Oriented-edge-only CPDAG

"Reliable causal discovery based on mutual information supremum principle for finite datasets". Cabeli, Li, Ribeiro-Dantas, Simon and Isambert. WHY21 at NeurIPS 2021.

Causal Discovery and iMIIC

Putative and genuine causal edges, and contextual variables



Contextual variables

"Genuine" vs "putative" causes

Reliability of orientations

Original MIIC

Causal Discovery and iMIIC

Genuine and putative edges in iMIIC

Endpoint head/tail orientation probabilities $X — Y \quad (p_t = 1 - p_h)$

- $p_{t_X} = 0.5, p_{h_Y} > 0.5$
  then $\longrightarrow$ **putative** cause $(\longrightarrow \; = \; \longrightarrow$ or $\leftarrow\!\!-\,-\!\!\rightarrow)$
- $p_{t_X} > 0.5, p_{h_Y} > 0.5$
  then $\longrightarrow$ **genuine** cause
- $p_{h_X} > 0.5, p_{h_Y} > 0.5$
  then $\leftarrow\!\!-\,-\!\!\rightarrow$ **latent** common cause $(\leftarrow\!\!-\,- L -\,-\!\!\rightarrow)$
- $p_{t_X} = 0.5, p_{h_Y} = 0.5$
  then $—$ **undetermined** or non-causal status

Endpoint head/tail orientation probabilities $X \text{ --- } Y \quad (p_t = 1 - p_h)$

- $p_{t_X} = 0.5, p_{h_Y} > 0.5$
  then $\longrightarrow$ **putative** cause $(\longrightarrow \; = \; \longrightarrow$ or $\longleftrightarrow)$
- $p_{t_X} > 0.5, p_{h_Y} > 0.5$
  then $\longrightarrow$ **genuine** cause
- $p_{h_X} > 0.5, p_{h_Y} > 0.5$
  then $\longleftrightarrow$ **latent** common cause $(\leftarrow L \rightarrow)$
- $p_{t_X} = 0.5, p_{h_Y} = 0.5$
  then --- **undetermined** or non-causal status

Prior knowledge about probability

- Contextual variable: $p_t = 1.0$

Causal Discovery and iMIIC

Genuine and putative edges in iMIIC

Endpoint head/tail orientation probabilities $X — Y \quad (p_t = 1 - p_h)$

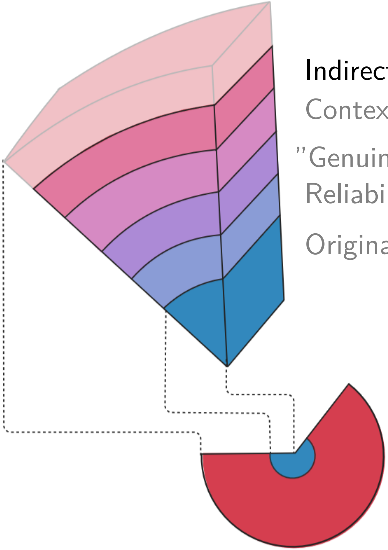
- $p_{t_X} = 0.5, p_{h_Y} > 0.5$
  then $\longrightarrow$ **putative** cause ($\longrightarrow = \longrightarrow$ or $\leftarrow\!\!-\,-\!\!\rightarrow$)
- $p_{t_X} > 0.5, p_{h_Y} > 0.5$
  then $\longrightarrow$ **genuine** cause
- $p_{h_X} > 0.5, p_{h_Y} > 0.5$
  then $\leftarrow\!\!-\,-\!\!\rightarrow$ **latent** common cause ($\leftarrow\!\!-\,-\,L\,-\,-\!\!\rightarrow$)
- $p_{t_X} = 0.5, p_{h_Y} = 0.5$
  then —— **undetermined** or non-causal status

Prior knowledge about probability

- Contextual variable: $p_t = 1.0$
  - Sex
  - YearOfBirth

Causal Discovery and iMIIC

Indirect path consistency



Indirect path consistency

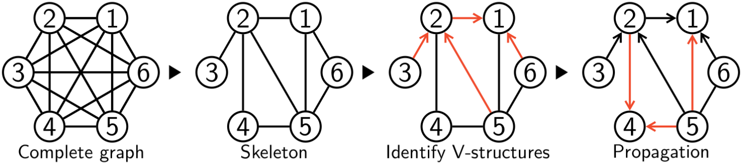
Contextual variables

"Genuine" vs "putative" causes

Reliability of orientations

Original MIIC

Motivation



Separating set: inconsistency

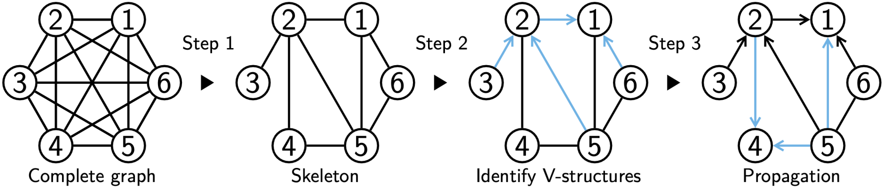
type I: $(2 \perp\!\!\!\perp 6 \mid 3)$ There is no path between 2 and 6 that goes through 3, inconsistent with respect to the skeleton;
type II: $(3 \perp\!\!\!\perp 6 \mid 1)$ The vertex 1 is a descendant of vertex 6 and 3, inconsistent with respect to the oriented graph.

In practice, these results, even if correct in terms of dependence relation, are **not interpretable**

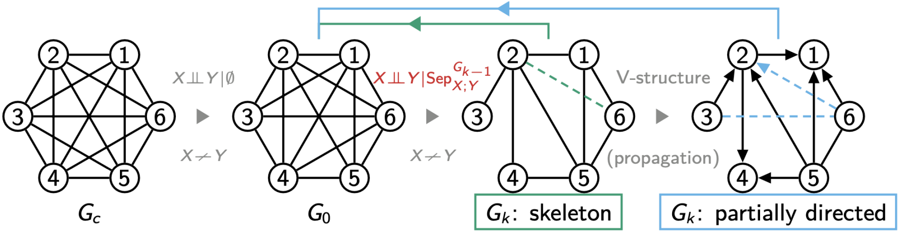
Causal Discovery and iMIIC

Consistency in iMIIC

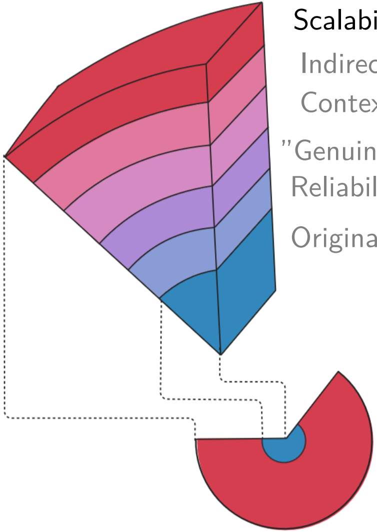
Classical Constraint-Based Methods present inconsistent separating sets!



$(2 \perp 6|3)$! $(3 \perp 6|1)$!



Li et al. Constraint-based Causal Structure Learning with Consistent Separating Set. Advances in Neural Information Processing Systems (NeurIPS) 32, 14257-14266 (2019).



Scalability & Performance

Indirect path consistency

Contextual variables

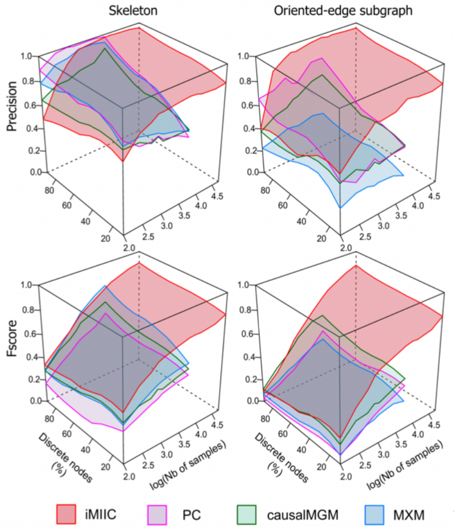
"Genuine" vs "putative" causes

Reliability of orientations

Original MIIC

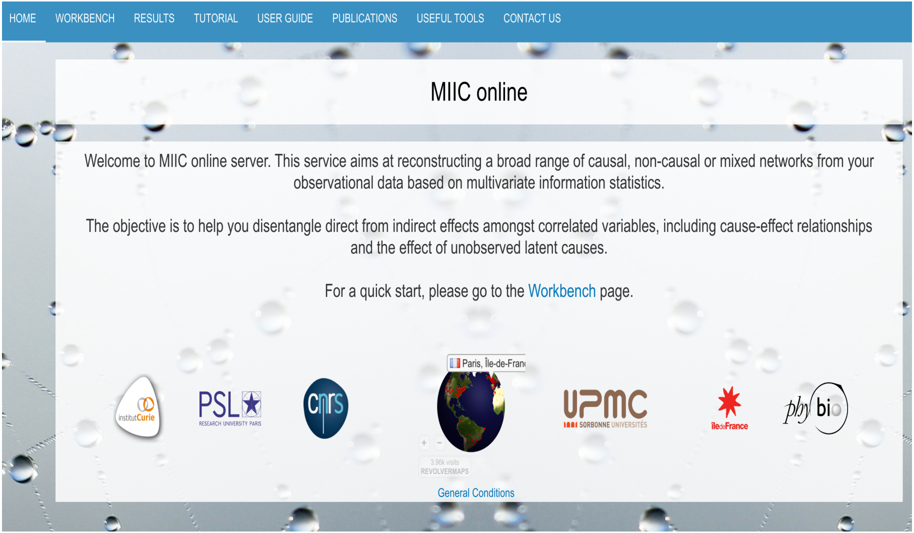
Causal Discovery and iMIIC

Scalability & Performance



MIIC WebServer

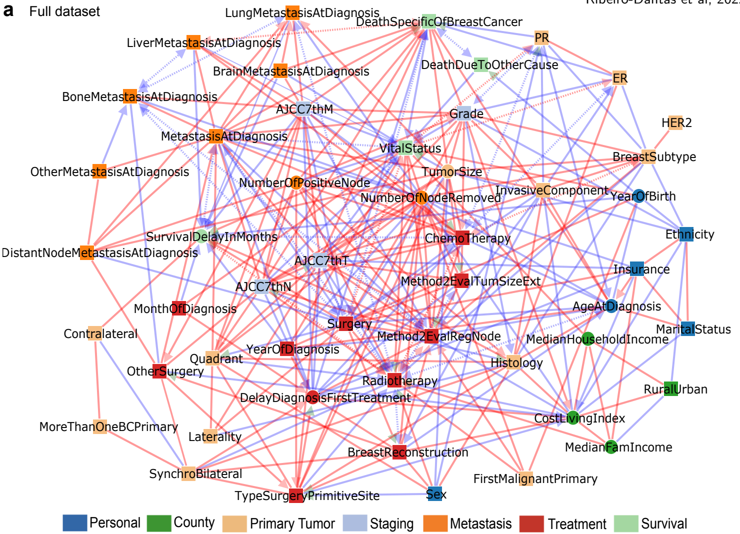
Demonstration



SEER network

Full skeleton-consistent network (396,179 samples)

Ribeiro-Dantas et al, 2022 (Submitted)

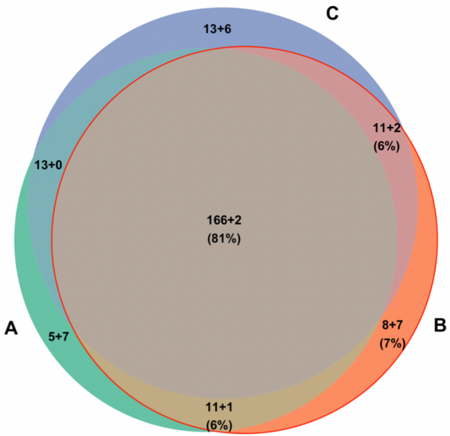


90% (resp. 98%) of causal edges are clearly correct (resp. plausible)

SEER network

Comparison of skeleton-consistent graphs of independent subsets of SEER

Overall robust inference with *few differences* (as many networks are 'nearly equivalent' for $N < \infty$)
3 independent 100k subsets ($a + b$ edges in intersections, $a \in$ full network, $b \notin$ full network)



88% of edge orientation probabilities are compatible bwn the three 100k networks
92% of those are also compatible with edge orientation probabilities of full network

SEER network: Analysis of network skeleton

Marital Status as Prognostic Factor

SCIENTIFIC REPORTS

Article | Open Access | Published: 31 January 2017

Prognostic value of marital status on stage at diagnosis in hepatocellular carcinoma

BJC British Journal of Cancer

Clinical Study | Open Access | Published: 17 May 2005

Sociodemographic factors and delays in the diagnosis of six cancers: analysis of data from the '*National Survey of NHS Patients: Cancer*'

SCIENTIFIC REPORTS

Article | Open Access | Published: 11 June 2018

Survival Comparisons Between Early Male and Female Breast Cancer Patients

PLOS | ONE   PUBLISH   ABOUT   BROWSE

OPEN ACCESS   PEER-REVIEWED

RESEARCH ARTICLE

Prognostic significance of marital status in breast cancer survival: A population-based study

María Elena Martínez, Jonathan T. Unkart, Li Tao, Candyce H. Kroenke, Richard Schwab, Ian Komenaka, Scarlett Lin Gomez

Published: May 5, 2017 • https://doi.org/10.1371/journal.pone.0175515

ELSEVIER

The Breast

Volume 32, April 2017, Pages 13-17

Original article

The effect of marital status on breast cancer-related outcomes in women under 65: A SEER database analysis

SEER network: Analysis of network skeleton

Marital Status as Prognostic Factor

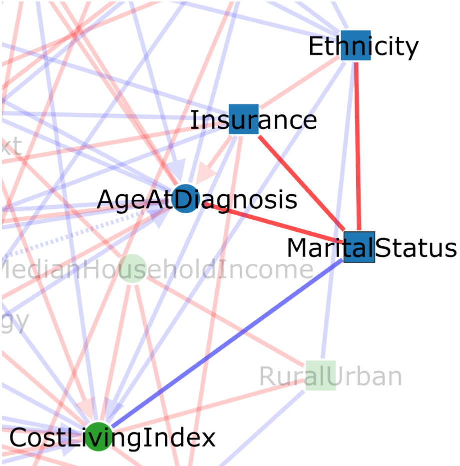
More information: Zhen Zhai et al, Effects of marital status on breast cancer survival by age, race, and hormone receptor status: A population-based Study, *Cancer Medicine* (2019). DOI: 10.1002/cam4.2352

"Our study demonstrates that patients with breast cancer could gain significant benefits from marriage and indicates the importance of psychosocial support to patients with unfavorable marriage," said co-author Zhijun Dai, of Zhejiang University, in China.

SEER network: Analysis of network skeleton

Marital Status as Prognostic Factor

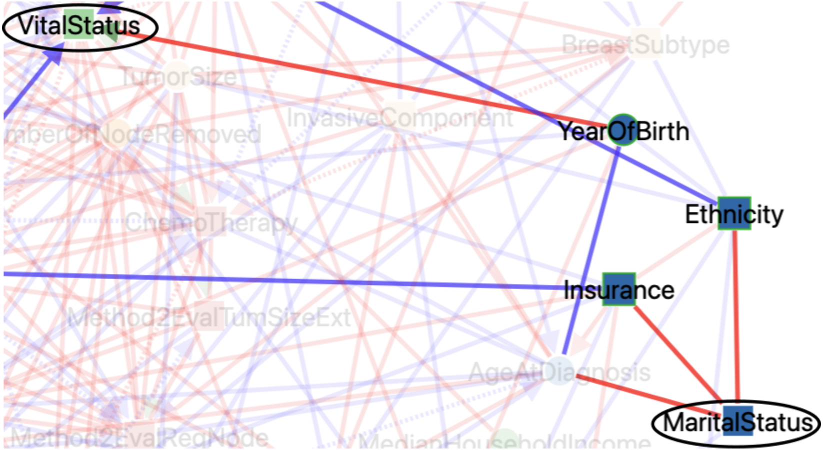
Marital status



SEER network: Analysis of network skeleton

Analyzing separating set

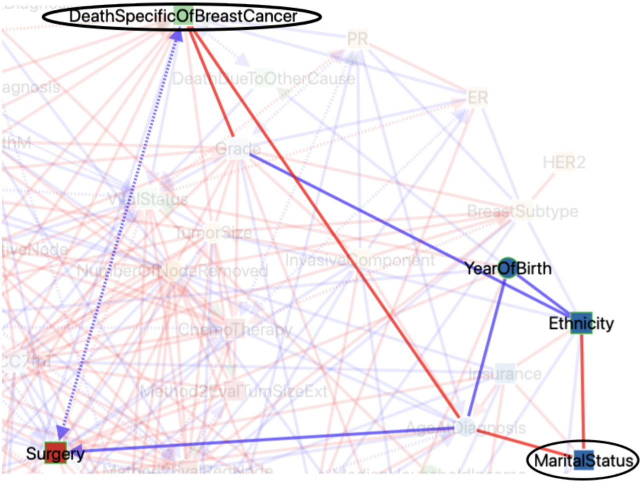
Why the edge between MaritalStatus and VitalStatus was removed?



SEER network: Analysis of network skeleton

Analyzing separating set

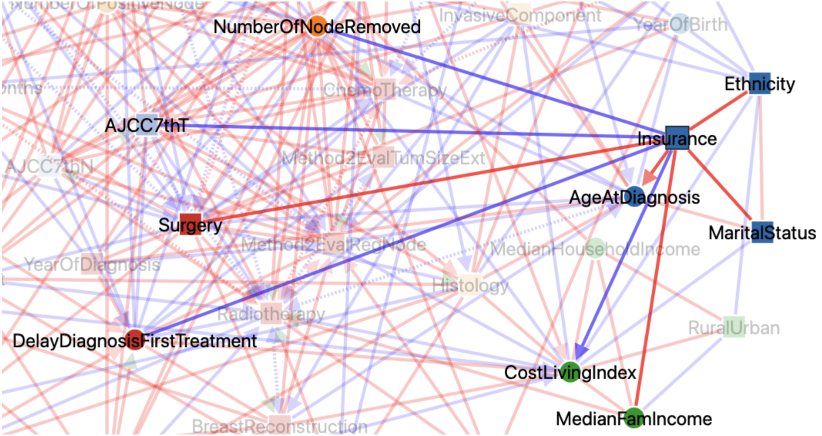
Why the edge between MaritalStatus and DeathSpecificOfBreastCancer was removed?



SEER network: Analysis of network skeleton

Analyzing separating set

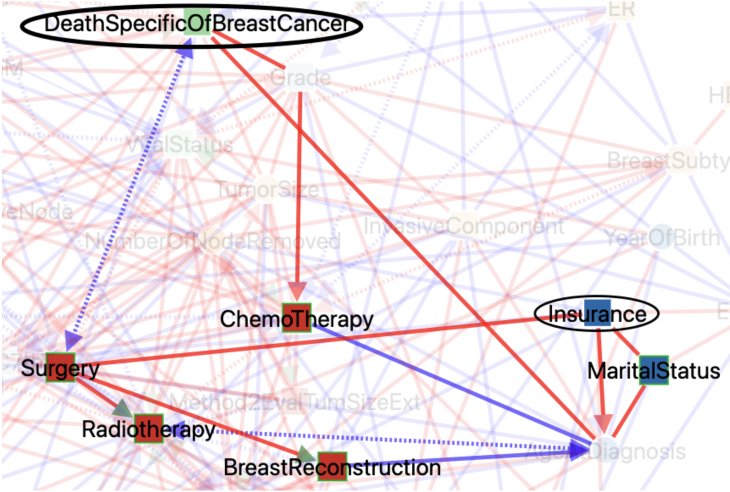
Insurance subnetwork



SEER network: Analysis of network skeleton

Analyzing separating set

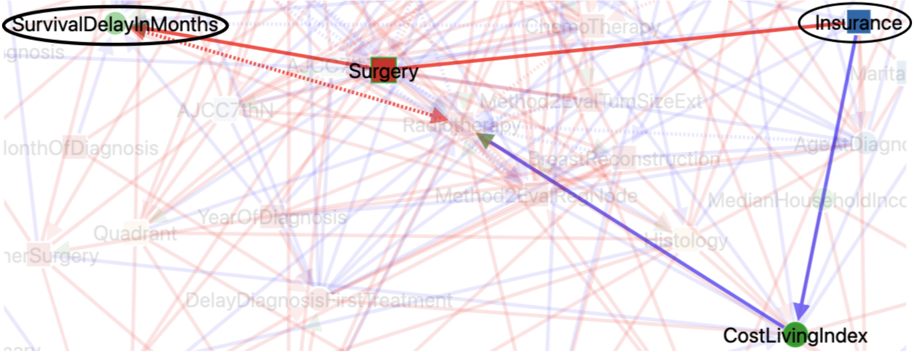
Why the edge between Insurance and DeathSpecificOfBreastCancer was removed?



SEER network: Analysis of network skeleton

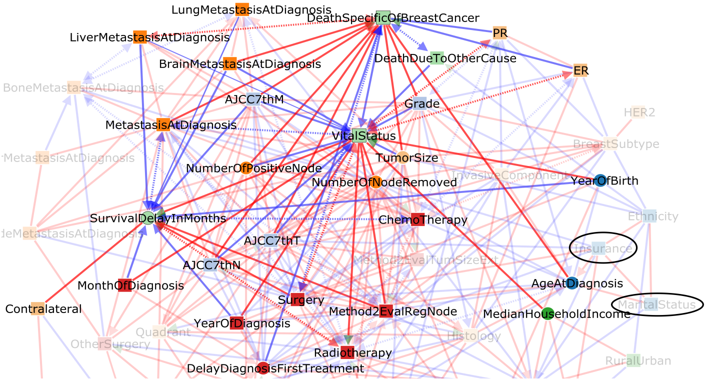
Analyzing separating set

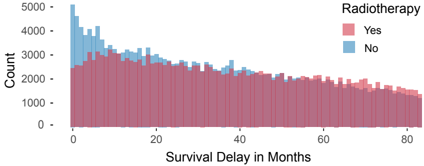
Why the edge between Insurance and SurvivalDelayInMonths was removed?

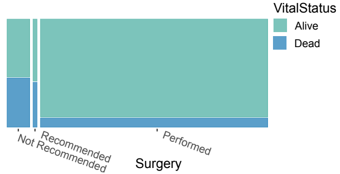


SEER network: Analysis of network

Survival subnetwork inferred by iMIIC from SEER breast cancer dataset

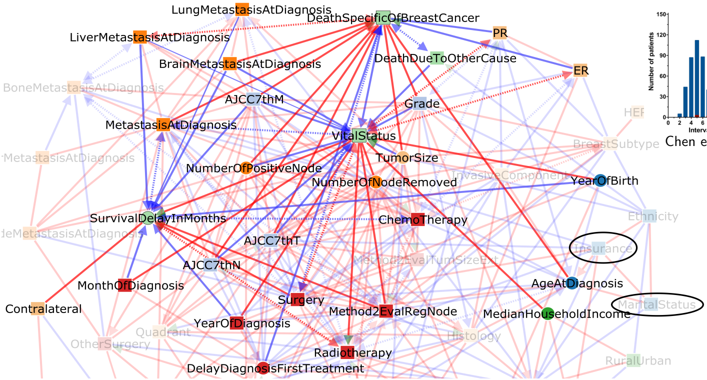


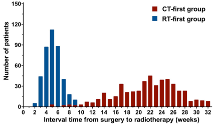




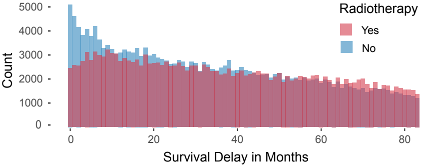
SEER network: Analysis of network

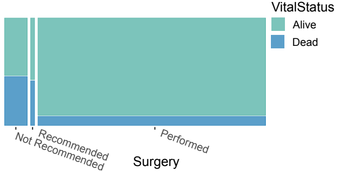
Survival subnetwork inferred by iMIIC from SEER breast cancer dataset





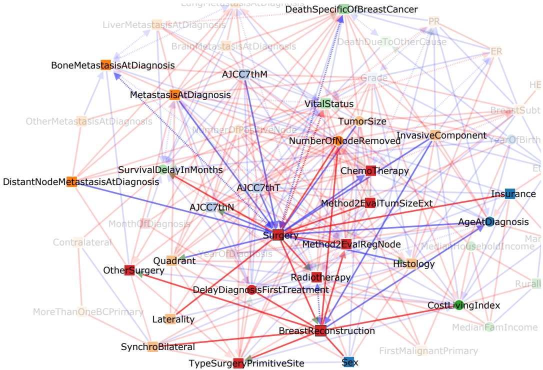
Chen et al. Front Onco 2020

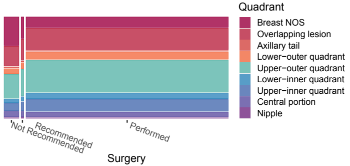


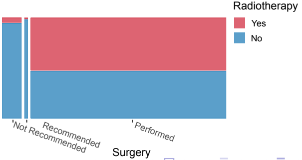


SEER network: Analysis of network

Surgery and subsequent treatments subnetwork inferred by iMIIC from SEER breast cancer dataset

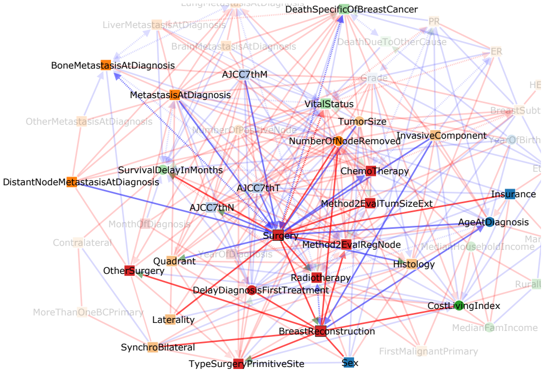




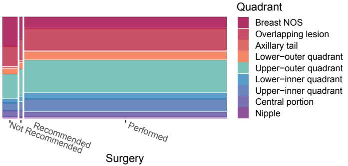


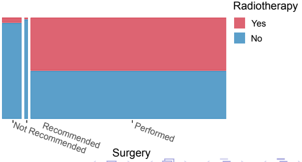
SEER network: Analysis of network

Surgery and subsequent treatments subnetwork inferred by iMIIC from SEER breast cancer dataset



Surgery ——▶ **Histology**
Infiltrating duct mixed with other types of carcinoma (+77% after surgery), Infiltrating duct and lobular carcinoma (+48%), Mucinous adenocarcinoma (+19%), Infiltrating duct carcinoma, NOS (+ 7.6%), Lobular carcinoma, NOS (-11%), Carcinoma, NOS (-91%), Adenocarcinoma, NOS (-95%)

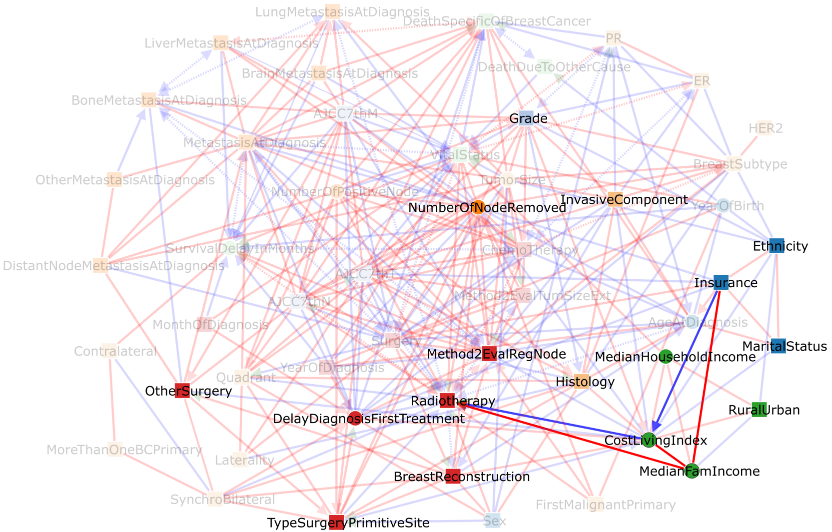




SEER network: Analysis of network edge orientations

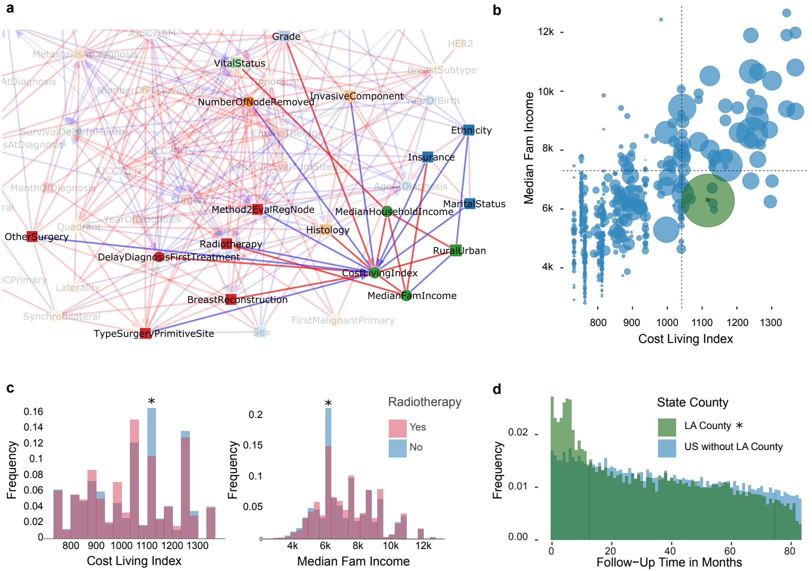
Analyzing genuine causal effects from CostOfLiving

The bigger picture



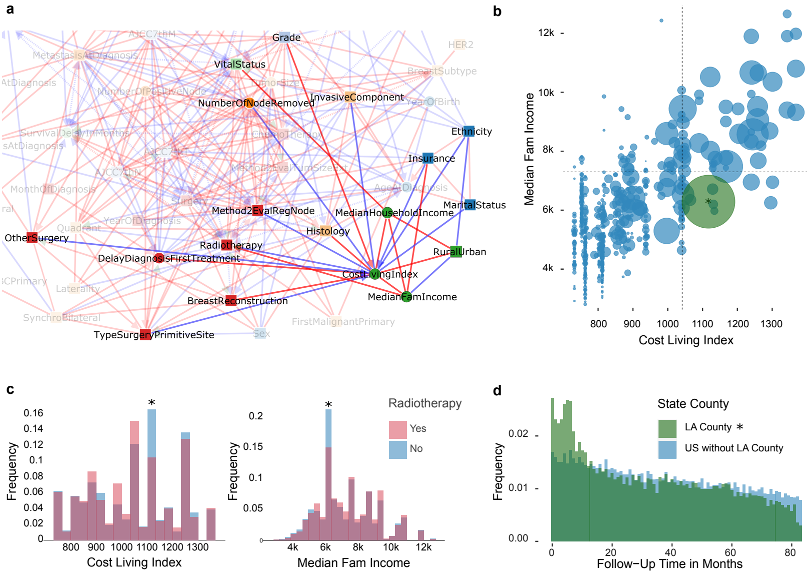
SEER network: Analysis of network

Socio-economic subnetwork inferred by iMIIC from SEER breast cancer dataset



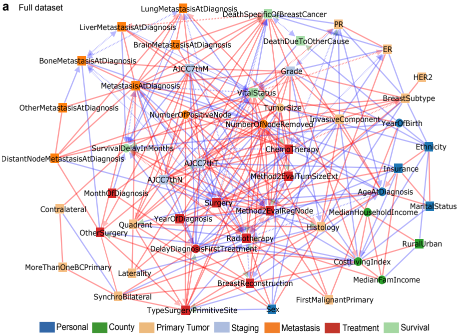
SEER network: Analysis of network

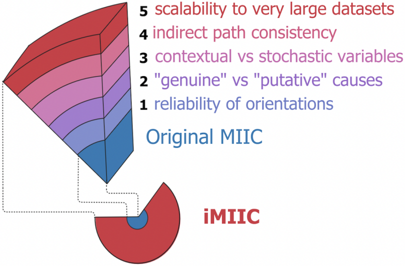
Socio-economic subnetwork inferred by iMIIC from SEER breast cancer dataset



∗L.A. county (10% of dataset): 30% Radiotherapy in L.A. vs 52% rest of US; 7% of L.A. patients drop out vs 1.5% rest of US

Key takeaways



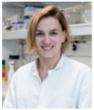


Acknowledgement


H. Isambert, PhD


V. Cabeli, PhD


F. Simon


AS Hamy, MD PhD


H. Li, PhD


N. Sella, PhD


L. Dupuis


L. Hettal, PhD

Commission européenne
MARIE CURIE ACTIONS
institut Curie
SORBONNE UNIVERSITÉ

Supplementary Materials

Causal Discovery and iMIIC

Consistency in iMIIC with consensus graph