# Learning interpretable causal networks from very large datasets, application to 400,000 medical records of breast cancer patients

Marcel da Câmara Ribeiro-Dantas

Sorbonne Université
ED130 – Ecole Doctorale Informatique, Télécommunications et Electronique

# Learning interpretable causal networks from very large datasets, application to 400,000 medical records of breast cancer patients

par Marcel DA CÂMARA RIBEIRO-DANTAS

Thèse présentée et soutenue publiquement le 4 Juillet 2022
en vue de l'obtention du grade de
docteur de L'Université Sorbonne Université
sous la direction de Hervé ISAMBERT

Membres du jury:

| | |
|---|---|
| Directeur: Dr Hervé ISAMBERT | Institut Curie |
| Co-Directrice: Dre Anne-Sophie HAMY-PETIT | Université de Paris |
| Rapporteur: Dr Jean-Christophe THALABARD | Université de Paris |
| Rapporteuse: Dre Laura CANTINI | IBENS |
| Examinatrice: Dre Michèle SEBAG | LRI |
| Examinatrice: Dre Nathalie VIALANEIX | INRAE |

# Acknowledgements

There are clearly many individuals and institutions who have helped me along the way on this journey, and it would not be fair to say this section is exhaustive. I apologize in advance for people I may forget to mention, but by no means should anyone think that I'm not thankful because someone is not mentioned here.

I am deeply grateful to all my professors, supervisors, and colleagues back in Brazil who helped me build a career that led me to be selected for this adventure in France, through the IC3i initiative, funded by Institut Curie and the European Commission as part of the Marie Curie Actions and Horizon 2020 funding programs. Such European institutions provided me with many of the best years of my life, intellectually and academically speaking and I couldn't be more grateful for that. I left Brazil a few years ago with a dream to become a better version of myself, to finally focus on research, in contrast to my time in Brazil when I had done research part-time, and I believe that at this point I have achieved this goal. Even though being selected for this position was clearly a great opportunity, I can't say the decision to come was an easy one, and for that, I thank everyone who supported me and believed in me, mostly my master's degree supervisor, Rodrigo Dalmolin, my bachelor's degree supervisor, Ricardo Valentim, friends, family and above all, my love, Natália.

Marci, Anna, and Aafrin, who came to Curie at the same time I did, were priceless companions on this journey. From the very beginning, when the three of us lived in the same building, the Collège Franco-Britannique in the Cité Internationale Universitaire de Paris, until the very end when the four of us lived far from each other, we still managed to meet and celebrate our friendship. It's easy for me to know who are the friends from my Ph.D. time that I will definitely never forget and that's Marci, Anna, and Aafrin.

Also, I would like to express my sincere gratitude to Hervé Isambert and Anne-Sophie Hamy-Petit, my Ph.D. supervisor and co-supervisor, who assisted me during this Ph.D. and contributed to shaping the researcher I have become. As part of the

team culture, Pr Isambert would have lunch with us every day, which certainly contributed to making my daily stay in Paris less lonely. Nadir Sella, Vincent Cabeli, and Honghao Li, the three Ph.D. students on the team at the time of my arrival, were amazing individuals, both as colleagues and researchers. I learned a lot from them, and I am thankful to them for every time we managed to do something together, both in the lab but also outside. Having lunch at Le Jardin du Luxembourg, learning to cook french food at Vincent's house, a barbecue at Nadir's, or playing games at Honghao's house. These were all very nice experiences and for all the support and friendship that you showed to me during this Ph.D., I'll never forget. For a short period Leonardo Renné and Maria Comes joined the lab and more recently, Franck Simon, Liza Hettal, and Louise Dupuis joined the team and I couldn't be happier about that. They not only added to the skills of the team, but are also great people! Spending my last months with them at Curie has been a joy and I'm thankful for every moment of assistance, and coffee times. French bureaucracy can be too much sometimes and Louise, Franck, and Liza helped me in several situations of bureaucratic despair haha When I thought Louise was already helping me too much, she would surprise me. Thank you!

I would like to offer my special thanks to the jury members and reviewers who agreed to participate in my jury and contribute to this work: Dr. Laura Cantini, Dr. Jean-Christophe Thalabard, Dr. Michèle Sebag, and Dr. Nathalie Vialaneix. I'm thankful to the administrative staff of the Physico-chimie Curie lab (UMR168) that have always been very helpful, and together with the whole unit, be it during retreats, conferences, or coffee time, made my time in Curie amazing and very happy.

For many reasons, I believe starting to work on causality in 2018 was great. Many recent discoveries contributed to progress in causal discovery, in many different fields, and a great book for beginners in causality was published by Judea Pearl, The Book of Why and The New Science of Cause and Effect. It was delightful to be introduced to causal inference by this book, and it definitely would have been harsher without it. For that, I thank Dr. Judea Pearl and all other experts in the field, including my Ph.D. supervisor Dr. Hervé Isambert, who gave me this wonderful opportunity to work with causal discovery in non-experimental data.

Lockdown surely was tough, but my childhood friends in Brazil, we refer to ourselves as the KND (Kids Next Door), didn't let me down. Many times we did video calls, played online silly games, made sure to keep everyone up in a nice mood. They've always been there for me, in over 20 years, and it was no different during these long years in France. Cadu, Cleto, Felipe, Ian, Daniel, Dr. Luciano Neto, and Harturo, thank you!

I’m thankful to my family who assisted me in any way they could during this period and throuhout my life. During my visits to Brazil, they made their best to make every second worth it, and never lost any hope in the great things they believe me to be able to achieve. For that, I’m very thankful to my father, Igor, my siblings Rafaella and Eric, and their partners Luiz Felipe and Manuela along with Eric and Manuela’s son, the amazing Bernardo! I would like to thank also my extended family, Natália’s parents, Ricardo and Tamara, and siblings, Nádia and Nicole. They’ve also supported me all the way here and never lost an opportunity to celebrate my visits during the Ph.D., and along with my family cheered for every progress I achieved.

Last, but not least, I couldn’t finish this section without thanking the person who gives me the strength to fight every battle, who pushes me forward every day, and who is at the same time my biggest idol and my biggest fan: My dear love, Natália. The willingness to make you happier and to work for a brighter future for us was primal in many of the difficult times that I had to go through in the past years. It’s no secret that a Ph.D. is a difficult journey to pursue, and having been physically alone throughout a large part of this experience, in another country, during the COVID-19 pandemic, among other things, contributed to what certainly was a more difficult version of what it could have had been in different circumstances. And yet, every day was a brighter day to me because I knew it was closer to the day I’d be back in your arms. Some people are afraid of endings due to the anxiety of the new, the next journey, the next day. I look forward to the next journey because I know that any journey will be delightful if I have you by my side. I love you, and for your love and everything that comes with it, including your support, I’m all grateful!

Having said all this, I finish this section by thanking again everyone who direct or indirectly contributed to my arrival at this place now, even if not explicitly mentioned here.

# Abstract

Uncovering cause-effect relationships in non-experimental settings has shown to be a very complex endeavour, given the numerous limitations and biases found in observational data. At the same time, there are many situations in which experiments can not be performed, be it due to technical, financial or ethical reasons, and large amounts of observational data are available. Recent progress in causal discovery methodologies, and in the causal inference literature in general, has contributed to the development of techniques that learn the underlying causal structure of the events recorded through observational data, allowing us to perform causal discovery and inference in observational data. The approach improved and used in the studies that this thesis describes is based on novel information-theoretic methods to analyze information-rich clinical data from curated clinical records as well as medical consultation reports of breast cancer patients. While numerous methods have been developed to identify correlations in heterogeneous clinical records, a central challenge remains: to uncover unsuspected cause-effect relationships when clinical essays are impractical or costly if not unethical despite possible benefits for the patients' health and survival. In such settings, it is now considered a priority to guide clinical understanding and treatments by novel and innovative data analysis and computational methods. Apart from skin cancer, breast cancer is the most common cancer in women in the United States, and the second leading cause of cancer death among women overall and the leading cause of cancer death among Hispanic women. Yet, there are few efforts to analyze the large amount of observational data related to this disease from a causal perspective. By analyzing the dataset generated by the Surveillance, Epidemiology, and End Results (SEER) Program in the US with the approach developed through the years at Isambert lab, mainly the iMIIC algorithm, it was possible to infer from data of approximately 400,000 patients diagnosed with breast cancer between 2010 and 2016 a network that presents many putative and genuine causal relationships, supporting previous discoveries in the

literature but also shedding light for new discussions.
**Keywords:** Causal Discovery, Causality, Breast Cancer, SEER, Bias

# Contents


**1 Introduction** **1**

1.1 Scientific context . . . 1

1.1.1 Causality . . . 1

1.1.2 MIIC . . . 2

1.1.3 Surveillance, Epidemiology, and End Results program . . . 3

1.2 Contributions . . . 3

**2 Causality** **5**

2.1 Causality discussions through the ages . . . 5

2.1.1 Ancient Greece . . . 5

2.1.2 Middle Ages . . . 7

2.1.3 Modern philosophy . . . 7

2.1.4 Recent causal literature . . . 12

2.2 Causal inference through the ages . . . 14

2.2.1 The biblical story of Daniel . . . 14

2.2.2 Scurvy, citricity and vitamin C . . . 15

2.2.3 John Snow and Cholera . . . 16

2.2.4 Smoking and Cancer . . . 18

2.2.5 Ladder of Causation . . . 20

2.3 Statistical Dependence . . . 22

2.3.1 Confounding bias . . . 23

Confounding by Indication . . . 24

2.3.2 Collider bias . . . 25

2.3.3 Correlation measures . . . 26

Partial correlation . . . 27

2.3.4 Information Theory . . . 28

Entropy . . . 28
Differential entropy . . . 31
Cross entropy . . . 31
Joint entropy . . . 32
Mutual Information . . . 33
Conditional Mutual Information . . . 34
2.4 Structural and Graphical Causal Model . . . 35
2.4.1 Graph Theory . . . 36
2.5 Network Inference and Causal Discovery . . . 38
2.5.1 Assumptions for Causal Discovery . . . 38
2.5.2 Simple approaches for network inference . . . 41
Covariance Matrix . . . 41
2.5.3 Constraint-based methods . . . 43
The PC algorithm . . . 43
MXM . . . 45
2.5.4 Score-based methods . . . 45
GES . . . 45
2.5.5 Other approaches . . . 46
LiNGAM . . . 46
CausalMGM . . . 46
2.6 Paradoxes and fallacies related to causality . . . 47
2.6.1 Berkson's Paradox . . . 47
Low-birth weight paradox . . . 47
Monty Hall Problem . . . 49
Sackett's study . . . 50
2.6.2 Simpson's Paradox . . . 51
2.6.3 Table 2 Fallacy . . . 53

**3 iMIIC and iMIIC WebServer 55**
3.1 Network reconstruction . . . 56
3.1.1 Consistency for Separation Sets . . . 59
3.1.2 Latent confounder, putative and genuine orientations . . . 60
3.2 iMIIC WebServer . . . 61

**4 SEER 65**
4.1 SEER Program . . . 65
4.2 SEER database . . . 66
4.3 Preprocessing . . . 66

**5 Learning interpretable causal networks from very large datasets, application to 400,000 medical records of breast cancer patients** **69**

**6 Reliable causal discovery based on mutual information supremum principle for finite dataset** **97**

**7 Conclusion** **109**

7.1 Interpretable Causal Discovery in Breast Cancer records . . . . . 109

7.2 Future Research . . . . . 110

**A Résumé long en français** **111**

**References** **129**

# List of Tables

2.1 Age and comorbidity of patients aged $\geq$ 75 years who received surgery or PET for breast cancer in the south of the Netherlands between 2001 and 2008 . . . . . 25

# List of Figures


2.1 Ladder of Causation . . . . . 21

2.2 Overall survival of patients aged $\geq$ 75 years who received primary endocrine treatment in the period 2001–2008 in the south of the Netherlands (n=184) vs. all patients aged $\geq$ 75 years treated with primary surgery in the same region and time period (n=1504) . . . . . 24

2.3 Entropy of a fair coin. . . . . 29

2.4 Venn Diagram. . . . . 33

2.5 Venn Diagram. . . . . 35

2.6 This structure is known as fork. $Z$ is a confounder. If you don't consider/measure $Z$, you will observe some statistical dependence between $X$ and $Y$ but there will be none once you take $Z$ into consideration, that is, $X \not\perp\!\!\!\perp Y$, but $X \perp\!\!\!\perp Y|Z$. . . . . 40

2.7 This structure is known as chain. $Z$ is a mediator. If you don't consider/measure $Z$, you will observe some statistical dependence between $X$ and $Y$ but there will be none once you take $Z$ into consideration, that is, $X \not\perp\!\!\!\perp Y$, but $X \perp\!\!\!\perp Y|Z$. It's important to notice that the same thing would have happened if we had instead $X \leftarrow Z \leftarrow Y$. . . 40

2.8 This structure is known as V-structure. $Z$ is a collider. If you don't consider/measure $Z$, you will observe no statistical dependence between $X$ and $Y$ but there will be some dependence once you take $Z$ into consideration, that is, $X \perp\!\!\!\perp Y$, but $X \not\perp\!\!\!\perp Y|Z$. . . . . 40

2.9 PC Algorithm . . . . . 44

2.10 Smoking during pregnancy has a direct causal effect on mortality of child and an indirect effect through the effect on birth weight, i.e. Smoking → Mortality of Child, and Smoking → Birth weight → Mortality of Child. On the other hand, smoking during pregnancy is not the only cause for low birth weight. There are more severe causes for this that not only have a direct effect on mortality but also an indirect one through birth weight. These unmeasured other causes are named U here and one example is Congenital Brain Injury. It's possible to see that birth weight is a collider and if you only look at babies with low birth weight, you're conditioning on this node and therefore adding spurious dependence to your estimate. . . . . 48

2.11 There are three doors. In this diagram, it's shown the door you chose, the door Monty Hall opens and the door in which the prize is located. It makes no sense for Monty Hall to open the door you chose, or the door where the prize is hidden. So there is an arrow from your first door and the prize door directing him to the only door he can open. This means, that even though the door you chose and the door with the prize are independent, when you know which door he opened (you adjust for a collider), there is some spurious correlation between your door and the prize door and that's why you have a higher chance of changing. . . . . 49

2.12 monty hall . . . . . 50

2.13 Even if you understand that lack of fuel or a discharged battery can lead a car engine not to start (only one of them, or both together), knowing information about one of them doesn't help you predict the other. If you check the fuel tank of a car and I ask you to predict if the battery is charged or not, there is no information to help you. You're just clueless. Why? Because fuel tank and battery being charged are independent. But if you checked the fuel tank and there is fuel, and I tried to turn on the car and it didn't start, you now know that the battery is discharged because it's the only possibility. Knowing one, and adjusting for a collider (the car doesn't start) explains away the possibility of being lack of fuel: It has to be a discharged battery! . . 50

2.14 By only looking at hospitalized patients, which means adjusting for a collider, Sackett observed spurious correlation between the two disease groups. This is another example of the explain away effect. Once he looked at hospitalized and non hospitalized patients, was not adjusting for a collider anymore, he found independence, which is indeed the correct estimate from the true model. . . . . 50
2.15 When adjusting for stone size, we see treatment A being better than treatment B. . . . . 51

3.1 The graph on the left shows a point in a time in which there was a path between $A$ and $D$ through $C$ and $A \perp\!\!\!\perp D|\{C, F\}$. However, in the graph on the right, which is the final skeleton, it's possible to see that at some later moment the edge between $C$ and $D$ was removed, therefore $C$ is not in a path between $A$ and $D$ anymore, which makes the separation set $\{C, F\}$ inconsistent with the final graph. . . . . 59
3.2 At some point, the edge between $C$ and $D$ was removed because $C \perp\!\!\!\perp D|B$. However, in the final graph it is seen that $B$ was not a mediator, but actually a collider, a common descent of $C$ and $D$ and therefore it should not be considered for the separation set. The separation set $\{B\}$ is thus inconsistent to the final oriented graph. . . . . 59
3.3 Summary statistics retrieved from https://miic.curie.fr/job_results.php?id=SEER2022. There are more statistics scrolling the window to the right. . . . . 62
3.4 Triplets statistics retrieved from https://miic.curie.fr/job_results.php?id=SEER2022. There are more statistics scrolling the window to the right. . . . . 63
3.5 Data dictionary retrieved from https://miic.curie.fr/job_results.php?id=SEER2022 . . . . . 63
3.6 Job comparison retrieved from https://miic.curie.fr/ . . . 64

# List of abbreviations

| Abbreviation | Label |
|---|---|
| ATE | Average Treatment Effect |
| AIC | Akaike Information Criterion |
| BIC | Bayesian Information Criterion |
| CPDAG | Completed Partially Directed Acyclic Graph |
| DAG | Directed Acyclic Graph |
| DCCPS | Division of Cancer Control and Population Sciences |
| ECA | Exploratory Causal Analysis |
| GES | Greedy Equivalence Search |
| ICA | Independent Component Analysis |
| iMIIC | interpretable MIIC |
| KL | Kullback-Leibler |
| MAG | Maximal Ancestry Graph |
| MIIC | Multivariate Information-based Inductive Causation |
| NML | Normalized Maximum Likelihood |
| PAG | Partially Ancestry Graph |
| PC | Peter & Clark |
| PCC | Pearson's Correlation Coefficient |
| RSP | Surveillance Research Program |
| SCM | Structural Causal Model |
| SEER | Surveillance, Epidemiology, and End Results |

Chapter 1

# Introduction

"I would rather discover one causal law than be King of Persia."
Democritus (460–370 B.C.)

## 1.1 Scientific context

### 1.1.1 Causality

In the field of Artificial Intelligence, it's very common to discuss causality in terms of reasoning, on how organisms learn to differentiate between correlation and causation. Judea Pearl, in his famous The Book of Why: The New Science of Cause and Effect mentions that most animals, as well as learning machines, learn from associations, which is mere statistical correlation (Pearl and Mackenzie, 2018). Tool users, such as early humans and the *Corvus moneduloides*, a species of crow, act by planning and not simply imitation (von Bayern et al., 2018). To some degree, they not only plan an intervention, perform it and think about what happened, but they can also imagine about what could have happened. Babies, for example, perform a lot of experiments and, presumably, according to Pearl, that's how they acquire much of their causal knowledge. Interestingly, Pearl believes that by building a causal inference engine, we can make learning machines understand causality, just like some natural organisms can (Pearl and Mackenzie, 2018). A deeper discussion on what causality means today, and how this definition evolved through time, is presented in chapter 2.

### 1.1.2 MIIC

Differently from what trialists do, that is, to perform experiments to understand causal relationships (Piantadosi, 2017), there are many researchers in different fields such as statistics, epidemiology, economic sciences, among other, who study causal inference in non-experimental data, also known as observational data. Within this group of researchers, some of them, also referred as *Pearlians*, see benefits in using graphs to discuss and investigate causal relationships. There are many ways to create such diagrams, and this work is mostly focused on causal discovery, also known as exploratory causal analysis, which is the task to learn such graphs from data (Glymour et al., 2019).

The MIIC algorithm has been developed for many years now in Isambert lab at Institut Curie. It was based on the 3off2 scheme, a causal discovery algorithm for discrete datasets that provided a more robust approach to reconstruct graphical models from finite datasets, combining constraint-based and score-based approaches to infer structural independencies based on the ranking of their most likely contributing nodes (Affeldt and Isambert, 2016). As an application, authors applied the 3off2 scheme to reconstruct the hematopoiesis regulation network based on single cell expression data.

Later, MIIC was extended to work with unobserved latent variables (Verny et al., 2017) and the approach was applied to reconstruct different biological size and time scale networks, from gene regulation in single cells to whole genome duplication in tumor development as well as long term evolution of vertebrates. MIIC was then once again extended with several pre- and post-processing analyses allied to a web application that allowed users to easily upload their data, run the analyses and obtain a causal/non-causal network with several metrics and resources to investigate the inferred network (Sella et al., 2018). One example was shown with a regulatory network from 2167 single-cell gene expression profiles of blood stem cells, and another one with an inherently non-causal network corresponding to the physical contact map of amino acid residues within a protein structure reconstructed from 12,533 aligned homologous sequences of an abundant protein domain family: the response regulator receiver domain. Afterwards, more progress was made extending MIIC to work with continuous and mixed-type data (Cabeli et al., 2020). In the aforementioned study, MIIC was used to reconstruct a clinical network from the medical records of 1,628 elderly patients consulting for cognitive disorders at La Pitié-Salpêtrière hospital, Paris. And more recently, some efforts happened to bring separation set consistency to MIIC in a way that would allow consistency in the interpretation of the inferred networks (Li et al., 2019). More details about the

MIIC algorithm, including progress made with contributions by this thesis, will be presented in chapter 3.

### 1.1.3 Surveillance, Epidemiology, and End Results program

The Surveillance, Epidemiology, and End Results (SEER) program is supported by the Surveillance Research Program (RSP) in the Natural Cancer Institute's Division of Cancer Control and Population Sciences (DCCPS) to collect and publish data on individuals in the US diagnosed with cancer since 1973. The program registries in charge of collecting the data are currently in regions of the US that cover approximately 48% of the US population including about 42.0% of Whites, 44.7% of African Americans, 66.3% of Hispanics, 59.9% of American Indians and Alaska Natives, 70.7% of Asians, and 70.3% of Hawaiian/Pacific Islanders. The total number of variables in the program, depending on the dataset, can go over 600 and there are studies extending this number by merging the SEER datasets with other public datasets such as the Medicaid dataset (Warren et al., 2002). The set of all variables includes clinical data, socioeconomic data, personal data, bio-molecular data, among other, making it a very rich and heterogeneous dataset. The chapter 4 will present the SEER program and its data in more details.

## 1.2 Contributions

The main goal of this thesis is to allow MIIC to handle efficiently *large-scale* real-world datasets and to improve the interpretability of the causal network discovered by MIIC. This includes improvement of the MIIC algorithm, of the MIIC WebServer and the application of both softwares to the SEER dataset, mentioned in the previous subsection.

The online version of MIIC (Sella et al., 2018) has been partially rewritten, refactored and extended with more features, such as comparison of inferred networks. Besides, there were improvements in the consistency algorithm and not only on the way orientations are calculated, but also new orientation types were investigated, as described in chapter 3 and 6. With this new and latest version of MIIC, called iMIIC, we applied it to a subset of the SEER dataset, consisting of 51 mixed-type variables of 396,179 individuals diagnosed with breast cancer between 2010 e 2016 in the United States. This contribution will be shown in more detail in chapter 5 and is currently under consideration of a peer-reviewed international journal.

Chapter 2

# Causality

> "We do not have knowledge of a thing until we have grasped its why, that is to say, its cause."
> Aristotle (384−322 B.C.)

This chapter starts with an overview of the development of what constitutes causality through history (2.1 and 2.2), followed by how to quantitatively attempt to measure the statistical relationship between events (2.3). An introduction to Structural Causal Models follows (2.4), then paradoxes related to causality (2.6) and I end the chapter talking about Causal Discovery, also known as Exploratory Causal Analysis (ECA) (2.5).

## 2.1 Causality discussions through the ages

### 2.1.1 Ancient Greece

It is not an easy task to precisely pinpoint when discussions around causality first emerged. However, this historical perspective often starts with Ancient Greece and Socratic Philosophy, through Plato (424-348 B.C.) who apparently was the first one to attempt to put into words the principle of causality: "Everything that becomes or changes must do so owing to some cause; for nothing can come to be without a cause" (Plato, 2019). It was his student though, Aristotle (384−322 B.C.), who investigated further the metaphysics of causality, in what became known as the four types of explanation. In Posterior Analytics, he explains that knowing what a thing is implies knowing its *aitiai*, "something without which the thing would not be", and

that this could be explained in four different ways (Aristotle, 1994).

If we look at a marble statue of Apollo, an example given by him, we can answer "What is this?" in four different ways:

- This is marble.
- This is what was made by Phydias, the sculptor.
- This is something to be put in the temple of Apollo.
- This is Apollo.

Such explanations answer the following four questions: "What is this made of?", "Who is this made by?", "What is this made for?" and "What is it that makes this what it is and not something else?". These answers came to be known respectively as the material cause, the efficient cause, the final cause and the formal cause. Aristotle argued that the most important cause was the formal one, which is the one that, by the way, Plato was mostly interested in.

If we take into consideration common sense, when we think of causality we are usually referring to the efficient cause (or efficient aitia), what caused this Apollo marble statue to be, and the answer is that it was Phydias, its sculptor. Having read this, one can wonder: If everything happens because of a cause, it is reasonable to assume that there should be a first cause that was not caused by anything. Aristotle named this the "Unmoved Mover", *primum movens* in Latin, and is also known as the unchanged form, as Plato and Aristotle discussed causality in terms of matter, form and changes in such forms. For Aristotle, there is a mean and purpose for everything and this "why" is a property of everything, even if the thing have no consciousness (plants, stones, they all have purposes).

Stoic Philosophers brought some new cards to the table. By believing that everything is providentially ordained by fate, they stated that every event occurs necessarily due to certain causal conditions. They elaborated the idea that a cause is linked both to the concepts of regularity and necessity. All this reasoning was convenient for them, for if there was a single event that occurred without a cause, this would contradict their faith. According to Alexander, The Great (356−323 B.C.), for example, under the same circumstances, the same effect should be observed and it is not possible that it is not. According to such philosophers, the lack of perfection in such observations is due to our ignorance of the causal connections between events (Long, 1996). It's interesting to mention the ignorance of Greeks at the time when it comes to probability, mainly because most people today don't realize how recent the contemporary knowledge of statistics is. Gambling is very old, and back

in time Greeks had their dice-like item: It was made of animal ankle bones. Many philosophers were against gambling, and it actually makes a lot of sense, after all if everything is preordained, everything will happen according to the plans and it makes no sense to gamble. This sort of belief prevented great minds of the time to study what eventually became the field of probability. This dice-like object from animal ankles called *astragali* had 6 sides but only 4 sides allowed the dice to rest stable. One of the gambling games consisted of throwing it 4 times and the most valuable combination was, surprisingly, not the most uncommon one, though it wasn't very common either, which is evidence they had no knowledge about probability (Mlodinow, 2009).

### 2.1.2 Middle Ages

During the middle ages, some christian philosophers attempted to reconcile the views of Aristotle on causality with the Christian idea that God created the world out of nothing. They basically did it renaming the "Unmoved mover" of Aristotle to God, the creating cause of existence. By doing this, they disentangled causality in two types of efficient causes: *causa prima* and *causa secunda*, in which the former would be the originating source of being, and the latter everything else.

Thomas Aquinas (1225-1274) discussed causality, but not only that, to argue for, among other things, the existence of God. Even though he was a friar and a priest, such an attempt was not uncommon at the time. He used to see efficient causes as subordinate to final causes, defending that created things had as primary goal self-realization but this goal coincides / is subordinate to the final goal, which is God's intentions. Aquinas made a distinction between two kinds of efficient causes, *loose* causes and *tight* causes. Tight causes are known today as necessary causes, as in if A is a necessary cause for B and you see B happening, A necessarily happened. Loose causes are known today as sufficient causes in the sense that if A is a sufficient cause for B, and you observe B happening, A not necessarily happened, as other things may have caused B to happen.

### 2.1.3 Modern philosophy

Around the 17th century, there was a radical change in the way of thinking of cause and effect. It's when formal causation and final causation were rejected as reasonable forms of causation and only efficient causation is seen as proper explanation of causality, driven by René Descartes (1596-1650) that had brought a more mechanical view on causality. He left behind ideas of forms and transference of form

and the whole idea of 4 causes itself, and thought of it in a more physical way. Even though he acknowledged the concept of final causes, based on God's intentions, he saw no usefulness in investigating this. Not surprisingly, he still endorsed the two types of efficient causality, in which God would be the general cause, and particular causes would be the general principles or laws of nature (Miller, 1984).

Thomas Hobbes (1588-1679) was aligned to Descartes' views on causality, to the point of explaining even psychological and sociological phenomena in terms of causal relations between moving bodies. He defined causes as "the aggregate of accidents in the agent or agents, requisite for the production of the effect", and an effect as "that accident, which is generated in the patient" (Hobbes, 1839). But based on the fact that such accidents are bodies in movement, causation at this point is seen as something related to motion. Not the bodies or substances, but they when in movement are the causes. This idea of bodies and movement were so physical to Hobbes that he believed no action could occur at distance, without contiguous bodies, without physical contact.

Even though Spinoza (1632-1677) insisted on the two different types of causation, one for God (he called it genuine cause or free cause) and necessary causes for the rest, he rejected the final causation as a fiction, differently from other philosophers such as Descartes that saw it related to God (Miller, 1984). According to Spinoza, "This opinion alone would have been sufficient to keep the human race in darkness to all eternity if mathematics, which does not deal with ends but with the essences and properties of forms, had not placed before us another rule of truth" (Spinoza). Spinoza insisted on a more logical approach to causality, as cause and effect logically necessitating each other.

When it comes to Leibniz (1646-1716), he believed that there was nothing without a reason, and no effect without a cause, what can be seen as not only a defense for the existence of causal relationships but also of final causes. He rejected the idea of reducing the metaphysics of causality to motion (Descartes, Hobbes, and Spinoza). He agreed with the ideas of original causality, God, final causality and even saw final and efficient causality as complementary to each other. Each efficient cause happens in accordance with a general rule or final cause, which is preordained by God.

John Locke (1632-1704) held what is known today as a singularist approach to causation, which conflicts with the modern understanding of causality, ever since David Hume's contributions. He established a relationship between causation and power, but not to necessity. Locke said:

> Power being the source from whence all Action proceeds, the Substances

> wherein these Powers are, when they exert this Power into Act, are called Causes; and the Substances which thereupon are produced [...] are called Effects (Locke, 1847).

In his view, a cause is a substance putting its power to work, i.e. making an effect occur. When it comes to Isaac Newton (1642-1727), probably one the first things that comes to mind are his famous laws of motion, which are implicitly talking about effects of causes or lack thereof. When Newton wanted to make clear the difference between *true motion* and *relative motion* in his *scholium*, a marginal note or explanatory comment made by a scholiast, he explained what he meant by "cause":

> The causes by which true and relative motions are distinguished, one from the other, are the forces impressed upon bodies to generate motion. True motion is neither generated nor altered, but by some force impressed upon the body moved; but relative motion may be generated or altered without any force impressed upon the body. For it is sufficient only to impress some force on other bodes with which the former is compared, that by their giving way, that relation may be changed, in which the relative rest or motion of this body did consist. Again, true motion suffers always some change from any force impressed upon the moving body... (Newton, 1687).

The conflict regarding causality between Newton and other authors mentioned so far is that a body can be in movement with no cause, free, due to no force being currently applied to it. There is no law of universal causation for Newton, for events do not need necessarily a cause, and any movement that happens due to the first law of motion is a causeless event (Collingwood, 1937). In few words, for Newton, in the universe there are events that happen according to a law, and events that happen due to causes. At this point, we can already see how the way philosophers saw causality stopped evolving slowly, mostly with different names or interpretation for the same things, and started having radical changes.

David Hume (1711-76) stated that causal relations depend on three factors: contiguity (in space and time) between the case and the effect, priority in time of cause to effect, that is the future can not cause the past, and a necessary connection between the two. He used to give extra weight to the last one because it's usually how we as humans differentiate causal from non-causal relationships. The issue with this third factor is that there is no way to rationally justify it, and Hume knew this. The necessity that we imagine from causal relationships is illusory, born from our expectations, which are due to habit. The concept of regularity is what helps Hume

understand where this illusion comes from. Watching something happening once does not usually make us think of it happening due to a cause, it's easy to think of it just happening by chance, even in layman terms. Watching it happen in similar ways through time, however, suggests to us that there must be something causing this to happen in a similar way. This necessity of an existing connection, according to Hume, is projected onto the world by our minds, and not the opposite. Even today, the idea of regularity as a necessary condition for causation is generally accepted. However, it has been later shown that regularity as a sufficient condition for causation is false, as shown by Thomas Reid (1710-96). There are many examples of regularity or constant conjunctions that are not causal relations, such as day following night (Hulswit, 2004). Therefore, for Hume, causality was something in our minds. Maybe that's why David Lewis (1941-2001), a philosopher at Princeton, understood that Hume had given not only one definition, of regularity, but also a second one of counterfactuals, which is closer to the way we think about causality (Pearl and Mackenzie, 2018). If A causes B, had A not happened, and everything else happened the same way, B would not have occurred the same way. Actually, Lewis suggests we should abandon the regularity idea and interpret the causal relationship between A and B through counterfactuals, as exemplified above, (Lewis, 1973).

Kant's thoughts on causality were somehow influenced by what Hume had said before. Kant could not accept Hume's conclusion that neither causation nor induction could be rationally justified, which implies that we can not rationally justify scientific knowledge. At some point, Kant concluded that either there is no such a thing as causality, or it must be something grounded *a priori*, which can be seen as an anti-Humean conception of causality.

> If we thought to escape these toilsome enquiries by saying that experience continually presents examples of such regularity among experiences and so affords abundant opportunity of abstracting the concept of cause, and at the same time of verifying the objective validity of such a concept, we should be overlooking the fact that the concept of cause can never arise in this manner. It must either be grounded completely *a priori* in the understanding, or must be entirely given up as a mere phantom of the brain. For this concept makes strict demand that something, A, should be such that something else, B, follows from it necessarily and in accordance with an absolutely universal rule. Appearances do indeed present cases from which a rule can be obtained according to which something usually happens, but they never prove the sequence to be necessary. To the synthesis of cause and effect there belongs a dignity

> which cannot be empirically expressed, namely, that the effect not only succeeds upon the cause, but that it is posited through it and arises out of it. This strict universality of the rule is never a characteristic of empirical rules; they can acquire through induction only comparative universality, that is, extensive applicability. If we were to treat pure concepts of understanding as merely empirical products, we should be making a complete change in [the manner of] their employment (Kant, 1781).

Kant is not only in favor of the idea of necessity, which Hume had rejected, but also believes causality can not be established empirically.

> This may seem to contradict all that has hitherto been taught in regard to the procedure of our understanding. The accepted view is that only through the perception and comparison of events repeatedly following in a uniform manner upon preceding appearances are we enabled to discover a rule according to which certain events follow always upon certain appearances, and that this is the way in which we are first led to construct for ourselves the concept of cause. Now the concept, if thus formed, would be merely empirical, and the rule which it supplies, that everything which happens has a cause, would be as contingent as the experience upon which it is based. Since the universality and necessity of the rule would not be grounded a priori, but only on induction, they would be merely fictitious and without genuinely universal validity (Kant, 1781).

Kant therefore states that (a) every event has a cause; (b) the cause of every event is a prior event; (c) the effect follows from the cause necessarily, and (d) in accordance with an absolutely universal rule; (e) this is known to us not from experience but *a priori*.

John Stuart Mill (1806-73) held that what we usually refer to *the* cause is usually a partial cause, but because we want to draw attention to it, we call it *the* cause. According to him, given a set of conditions for something to happen, *the* cause is either the last condition before the effect happens, or the "superficially the most conspicuous" condition (Mill, 1874). He stated that this ordinary definition of cause is misleading and that it would be more appropriate to refer to cause as the set of conditions through which the effect invariably occurs.

> The cause, then, philosophically speaking, is the sum total of all the conditions, positive and negative taken together, the whole of the contingen-

> cies of every description, which being realized, the consequent invariably follows (Mill, 1874).

This makes much more sense, for it explicitly states the conditions in which the deterministic relationship between cause and effect will occur. A can cause B and yet the effect not be observed, if something else cancel the effect of A upon B. But he did not stop with invariability and insisted on unconditionality. He said:

> If there be any meaning which confessedly belongs to the term necessity, it is *unconditionalness*. That which is necessary, that which must be, means that which will be, whatever supposition we may make in regard to all other things. The succession of day and night evidently is not necessary in this sense. It is conditional on the occurrence of other antecedents. That which will be followed by a given consequent when, and only when, some third circumstance also exists, is not the cause, even though no case should ever have occurred in which the phenomenon took place without it (Mill, 1874).

The idea of unconditionally is a direct hit to empiricism, because it states that the reasoning not only depends on what we have seen, but also on what we have not seen but could happen.

Through history, even though the definition of cause and the set of ideas surrounding causality slowly changed, it's possible to separate these philosophers in two groups: The group that did not diverge much from what Aristotle had postulated, and the group that is closer to the view we currently have. It's easy to separate them because their views are mutually incompatible. The first group discussed the whole topic in a teleological way, that is, exhibiting or relating to design or purpose especially in nature, as if it was just following a plan, be this plan made by the christian God, Zeus or the universe itself. The second group could not find evidence of the foundations pointed out by the first group. They saw the causal relationship between entities merely following laws of nature.

### 2.1.4 Recent causal literature

More recently, efforts have started to link causality and statistics, more specifically through probability, with the idea that A causing B implies that the probability of B happening is increased by the occurrence of A. This started with Hans Reichenbach (1891-1953) and Patrick Suppes (1922-2014) (Pearl and Mackenzie, 2018).

After all the progress in these discussions through the ages, one may still look for evidence of the existence of causes and effects. Most equations that we see in the physics literature are symmetrical, in a way that cause and effect don't look so different, just like past and future don't look so different. We also have this idea that causes came before than the effects they caused, and still it is not trivial to prove this. Ludwig Boltzmann (1844-1906) contributed to the understanding of entropy in the second law of thermodynamics by viewing entropy as the number of ways that one can rearrange the constituents of a system. These discussions are related to the arrow of time, as irreversible processes are at the heart of the arrow of time and according to the second law of thermodynamics the total entropy in the universe can never decrease. If you start with low entropy, it's just natural that entropy will increase, according to Boltzmann, because there are more ways to be in a high entropy state than in a low entropy state, but what Boltzmann did not help clarify is why the universe was supposedly in a lower entropy state in the beginning, as supporters of the theory of past low-entropy of the universe. This topic is largely discussed by scientists such as Sean Carrol (Carroll and Chen, 2004).

Macroscopically, it's more intuitive to explain the difference between causes and effects. If you have all the ingredients for an explosion, you can cause an explosion or not depending on how you change these ingredients and conditions. They're the cause. On the other hand, when the explosion occurs you can not change something in the explosion in a way that it would change the past, such as the configuration of ingredients and conditions in that past. The explosion is the effect. If you captured a fish that was contaminated by industrial waste, it's intuitive to understand that this fish would not have been contaminated if there was no industrial waste, all other things kept constant. However, removing the contamination from this fish will not prevent industrial waste from happening to that body of water in the past. The contamination in the fish was the effect. This counterfactual view, thinking of a hypothetical world in which only the intervention intended was different, keeping everything else constant (also known as *ceteris paribus* in the economic sciences literature) is a very common approach to explain causality. If we could time travel, it would be trivial to identify causal relationships. Let's say you had a headache Saturday morning and took some water with some magical powder that you bought somewhere. By the end of the day, your headache was gone. Magical! Was it due to the water with magic powder? Let's go back in time, make you not put the magical powder in your water, keep everything else constant and see what happened in the future. If at night your headache was once again gone, we can say that the improvement was not caused by the magic powder. It would have happened anyway. One

common explanation for beliefs in alternate approaches has to do with delayed effect. An intervention A, such as a medicine, could take hours or even days to make you feel better. If you take an alternative solution (taking treatment B, for example) afterwards, closer to the effect of the first treatment, you may be more inclined to believe treatment B was responsible for your recovery, when actually it was the first treatment all along. There is also what is called the natural history of the disease. Disease A can show up with symptoms and go away by itself on a few days, without you having to do anything, but if you did something close to the end of it, you may be prone to believe it only went away because of what you did.

## 2.2 Causal inference through the ages

As we know, time travel (and also the keeping everything else constant detail) is not possible, so we have to find new ways to investigate causality. Through history, some people managed to do that and more recently all these strategies have been improved and systematized, but before talking about them, I will mention some events in history that were better understood through the attempt to understand causal relationships.

### 2.2.1 The biblical story of Daniel

According to Judea Pearl (Pearl and Mackenzie, 2018), one of the oldest if not the first controlled experiment to investigate causation happened with Daniel, according to a biblical story. After the king of Babylon had sacked the kingdom of Judea, he had taken captives with him back home. Among these captives were well educated youngsters such as Daniel. The idea was that such cunning children should be well fed and educated so that in the future they could serve in king Nebuchadnezzar's court, in the administration of the Babylonian Empire. The king demanded that these children should eat just like the king, but Daniel and some other children refused to eat those meals. Instead, he requested that they should be fed with a diet of vegetables. Asphenaz, the person king Nebuchadnezzar had left in charge of these kids, did not want to annoy the king, by not following his orders and therefore resisted to change the meals Daniel and the others were supposed to receive. Asphenaz was afraid that a change in diet in the way Daniel was requesting could make them look weaker or ill and he would be the one to suffer the ire of the king when the king noticed that. Attempting to convince him this would not happen, Daniel tried to reason with Asphenaz, with the following plan. Asphenaz should let them have the

diet Daniel requested for a few days, and at the same time, have another group of kids having the diet the king demanded. If after these days the kids in Daniel's group look fine, Asphenaz could rest assured that the diet wouldn't make them look less healthy. Even though it's easy to spot many limitations in this experiment, thinking of a comparison group, with similar characteristics, is a somewhat advance from just observing what happens in one group after an intervention. It's interesting to notice that even though this supposedly happened thousands of years ago, planned by a kid, we have adults right now in 2022 assuming causation from episodes that happened with a single friend. Something along the lines of: My friend felt bad, did this and that and he felt better. Then, clearly, what he did was the responsible for (the cause of) his improvement. Daniel knew better.

### 2.2.2 Scurvy, citricity and vitamin C

Scurvy was first described by Hippocrates and was known as "The Pirate's Disease" due to its prevalence among people who stayed for long periods of time on the high seas. Between the 15th and 17th century, due to the significant increase in sea travels and expeditions, millions of people died of scurvy. In one of his expeditions, Vasco da Gama lost 116 of his crew of 170; In 1520, Ferdinand Magellan, during the first circumnavigation of the world, lost 208 out of 230, mostly due to scurvy (Lamb, 2016). Based on so many cases and deaths related to scurvy throughout time, it's no surprise that many people tried to treat the disease, with many different rumors explaining how some sailors managed to cure it. Some people said they had been cured by drinking cider, or elixir of vitriol (a mixture of sulfuric acid and alcohol), or sea-water, or juice of citric fruits such as orange and lemons, and even by eating rats found in the ships!

James Lind (1716-1794), a Scottish physician working for the Navy between 1747 and 1753 got interested in this disease, and he decided to run an experiment, not much different from the one Daniel supposedly did a long time before. He took some men suffering from symptoms common to scurvy and separated them in pairs, giving a different treatment to each pair. They didn't listen to the rats rumors and focused on either (1) a quart of cider a day, or (2) 25 drops of elixir of vitriol, three times a day, or (3) half a pint of sea-water a day, or (4) a nutmeg-sized paste of garlic, mustard seed, horse-radish, balsam of Peru, and gum myrrh three times a day, (5) or two spoonfuls of vinegar, three times a day, or (6) two oranges and one lemon a day, a total of 6 different treatments. The story says that pretty soon the pair who had taken two oranges and one lemon, the citric fruits, were feeling so well that they assisted the treatment for the rest of the patients. Again, it's trivial to spot many lim-

itations in this experiment. We don't even know if they were followed exactly as the story says, and still we can think of lack of randomization, blinding, adequate sample size for the expected effect size and so on. Regardless of all that, Lind believed to have discovered that citricity was the cure for scurvy. And if citricity is the cure, the more citric a fruit, the best cure it is!

Lind retired shortly after the experiment, and even published a book on the topic (Lind, 1753). However, it was not until 1795 that the Royal Navy started introducing citrus rations, when Lind was already dead (Sutton, 2003). Brittish navy decided to distribute a lot of Indian limes in their ships, which was a great deal because Indian limes were not only more citric than lemons and oranges, but were also way cheaper. However, there is usually a negative correlation between citricity and amount of vitamin C, which means that the amount of vitamin C the sailors were receiving was not enough to really cure/prevent scurvy. It was only in 1919, many years later that Jack Drummond (1891-1952), an English biochemist, isolated the anti-scurvy factor nowadays known as vitamin C. The trivia here is that rats, among other animals but not human beings, are able to synthesize their own vitamin C, which is one reason for these animals not developing scurvy. Sailors who ate rats in the ships, to a certain degree, could have benefited by the vitamin C found in the animal, depending on how they prepared it. Maybe, experiments with rats would have shown earlier that citricity is not the cause behind the cure for scurvy!

### 2.2.3 John Snow and Cholera

In 1853 and 1854, there was a harsh cholera epidemic in England and, differently from today, it was as very lethal disease. A healthy person who drinks cholera-tainted water could die within twenty-four hours and, also very different from today, back in time they hadn't understood properly what was behind the disease. It's important to make it clear that even though at the time scientists were already aware of microorganisms, discussions on microorganisms causing disease in humans was still something recent. At the beginning of that century (1808-1813), the Italian entomologist Agostino Bassi had identified a germ that was responsible for devastating the French silk industry, and a bit later (1847) the Hungarian obstetrician Ignaz Semmelweis suggested "cadaveric particles" as responsible for the high death rate of women giving birth to doctors that had interacted with dead bodies right before. The germ theory of disease, thus, wasn't as accepted as it is nowadays. In the context of cholera, the miasma theory was the one usually mentioned to explain the cause behind the disease. Miasma refers to bad odors, but John Snow (1813-1858), an English physician, wasn't convinced. The symptoms of cholera, profuse watery

diarrhea, sometimes described as “rice-water stools” and vomiting, among other, were usually related to the intestinal tract and it didn’t really make sense for him that this could be caused by breathing some specific air. He suggested that whatever is causing the disease should interact with the intestinal tract of the individual, and not with the lungs.

John Snow thought of performing an experiment, but it should be clear by now how tricky performing such an experiment can be. Even if you ignore the ethics limitations of infecting someone with something that potentially causes a disease, how could he have groups of patients and infect one group if he didn’t even know what causes the disease? Besides, the distribution of some characteristics in these groups can have an effect on what we observe. Let’s say you’re testing a new treatment for a disease. You will give the new treatment to one group and regular treatment to the other group. If the group that gets the new treatment is at a very severe stage of the disease, and the group with a regular therapy is with a light version of the disease, even if the new treatment is better, it won’t look like so. Even if you had the same treatment for both groups, the group with weaker patients (more severe version of the disease) will die at a higher rate anyway. There are too many things to think of if you want to investigate causality adequately and in some situations it’s just not possible. The situation of John Snow looked like a lost cause, but he found out a way!

He noticed that districts supplied with water by Southwark and Vauxhall Company had a death rate eight times higher. Tainted water seemed more reasonable to him, since it does interact with the intestinal tract. However, one could say that the miasma was stronger in those districts. How would he even measure miasma!? However, there was something else that Snow noticed. There were districts with water supplied by both Southwark and Vauxhall Company, and Lambeth Company, and these districts still had a higher death rate in households being supplied by Southwark and Vauxhall Company. It’s the same district, one can’t claim it’s miasma in the district anymore. The interesting bit is that the houses supplied by these two companies did not differ in other characteristics. There was poor people, and rich people, getting water from one company and from the other. Small houses, and large houses. Randomized Controlled Trials, as a research design, was still in the future, but what Snow had was something close to it. He had the treatment (water from Southwark and Vauxhall) being randomly attributed to individuals in a population and the individuals getting this treatment did not differ in other ways from individuals getting another treatment (water from Lambeth Company). In Snow’s own words, he says: "No experiment could have been devised which would more thor-

oughly test the effect of water supply on the progress of cholera than this, which circumstances placed ready made before the observer. The experiment, too, was on the grandest scale. No fewer than three hundred thousand people of both sexes, of every age and occupation, and of every rank and station, from gentlefolks down to the very poor, were divided into two groups without their choice, and in most cases, without their knowledge."

This is currently known as a natural experiment, for obvious reasons. Snow also knew more things about this topic. He knew that Southwark and Vauxhall Company drew its water from the area of the London Bridge, downstream from London's sewers, while several years earlier Lambeth Company had moved its water intake upstream of the sewers. What happened, in the end, is that Southwark and Vauxhall Company customers were drinking water tainted by excrement of cholera victims. Unfortunately, just like happened with Lind, for a long time Snow hypothesis was ignored, but it was later proved to be correct.

### 2.2.4 Smoking and Cancer

Maybe one of the most commonly discussed topics about cause and effect in healthcare is the link between smoking and lung cancer. Yet, for many reasons, including ethical ones, nobody has ever conducted a randomized controlled trial to investigate this question. How do we know then that smoking causes cancer? For many decades this has been a very highly debated topic, with less debate against the link arising in recent years, but still new evidence every now and then is brought up to support the hypothesis of smoking causing lung cancer.

It's important to mention that until the 18th century, lung cancer was not even a well described disease. By 1900, there were less than 200 documented cases in the world. By 1950, however, it had become the most commonly diagnosed cancer in male individuals in the United States. One hypothesis is that it could had been diagnosed in the past as a different disease, but detailed autopsies in Germany suggest this was likely not the case (Proctor, 2012). What was the cause of this fast change in prevalence of the disease then?

In 1900, the per capita consumption of tobacco in the US was of 54 cigarettes per year, while in 1963 this number had risen to 4.345 cigarettes (Warner et al., 2014). This wasn't the only hypothesis, though. Some people believed the cause to be smoke from cars/industry, the influenza pandemic in 1918, and even a gene (or something else) that caused lung cancer and also increased the likelihood of someone becoming a regular smoker. Sir Ronald Fischer (1890-1962) was one of the proponents of this idea. To disprove Fischer's hypothesis, Jerome Cornfield came

with what was later known as Cornfield's inequality, the first formal method for sensitivity analysis in observational studies (Greenhouse, 2009). Sensitivity analysis is a technique to quantify how strong unobserved variables need to be in order to reasonably change an observed outcome (Cinelli et al., 2020a). Some assumptions are commonly taken into consideration when performing some analysis. Sensitivity analysis tackles the issue of not having measured, and adjusted for, the type of thing that Fischer was proposing. Cornfield was able to show that the strength of such unmeasured confounder had to be large to an extent that it would not be reasonable (Cornfield et al., 1959).

More recently, scientific progress allowed us to perform experiments to better understand the relationship between smoking and lung cancer. There is an approach called Mendelian Randomization, that takes advantage of nature just like Snow did in the cholera story (Emdin et al., 2017). Mendelian randomization uses genetic variants to determine whether an observational association between a risk factor and an outcome is consistent with a causal effect. We are born with our genome, the set of all genes in our cells organized in our DNA, and our choices and events throughout life do not change our DNA. Therefore, we can use some genes just like Snow used the water companies during his analysis of the cholera epidemic in London. With this in mind, scientists identified a gene called CHRNA5 that had different versions (alleles). Smokers who had a certain version of this gene, let's call it A, were more likely to smoke less when compared to smokers who had another version, let's call it B. When patients were grouped according to their CHRNA5 version, individuals with the version B, associated with heavy smoking, were more likely to die younger, due to cardiac disease or lung cancer. However, when scientists investigated non-smokers who had the version B, associated with heavy smoking, there was no change in life expectancy (Davies et al., 2018). Therefore, the observed life expectancy change was seen as due to smoking.

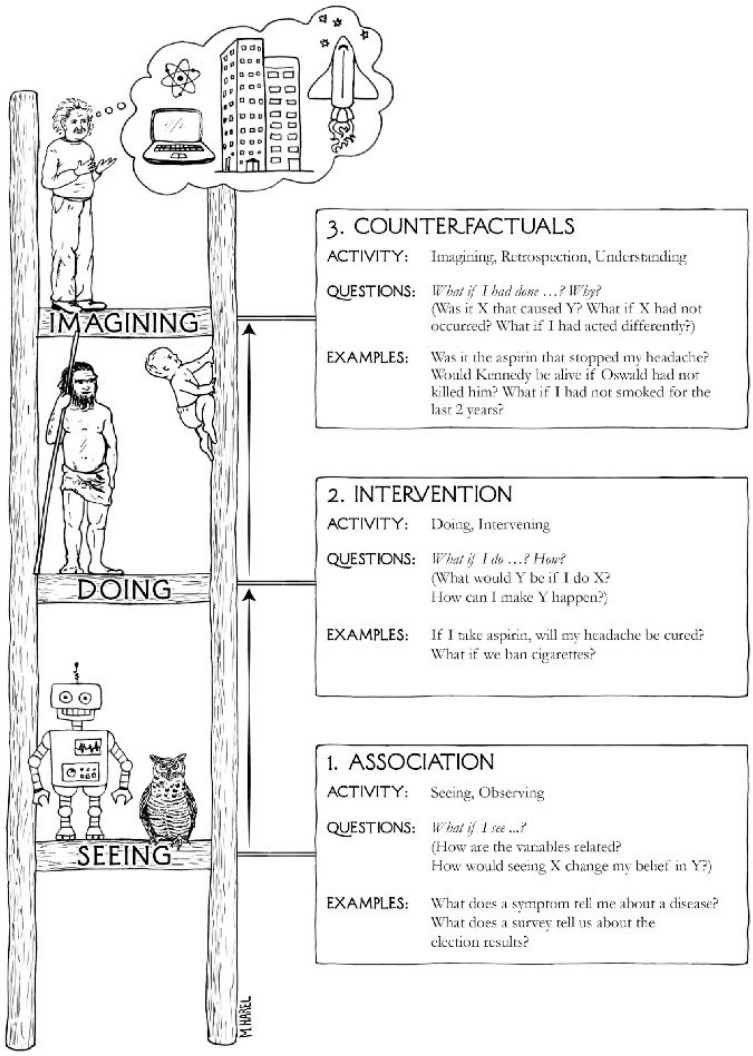
At the same time, in the discussion on smoking and lung cancer, Bradford Hill shared his viewpoints (later known as Hill's criteria for causation, contrary to how he named it at the time), as a qualitative approach to investigate causality. Some authors, such as Judea Pearl, acknowledge the progress in bringing a qualitative approach to the discussion, but believe such criteria to be no more than a historical document nowadays, to show us how causality was perceived in the recent past (Pearl and Mackenzie, 2018). The reason for that is that most of Hill's criteria can be easily debunked nowadays, with counterexamples. Strength (effect size), for example, has nothing to do with causality. You can measure tiny or huge statistical correlations, and this tells you nothing about causation. Consistency (reproducibil-

ity) doesn't really help either, as a badly designed study reproduced numerous times doesn't increase the likelihood that the effect observed is causal. It's a bit early in the text to really explain these limitations in detail, but I believe the next section will provide the tools and reasoning to better understand them.

### 2.2.5 Ladder of Causation

For over 20 years, authors have defined precisely a 3-level hierarchy for causal investigation Pearl (2000). Some authors conflate these levels, but making it clear they're different is important to understand causality and the limitation of some approaches. Though Figure 2.1a does a great work illustrating what they mean, to grasp its details, the next paragraphs will be useful.

Figure 2.1: Ladder of Causation




Retrieved from: (Pearl and Mackenzie, 2018)

Many animals and most current machine learning algorithms are at the first rung of the ladder. They learn by mere observation, and thus work based on association. One example of an association question is: If I see A happening, does that increase the chance of seeing B happening? There is no action or intervention here. You've been provided with some observation and you wonder about what you can see next or not. On the second rung we find early humans and a few other animals that managed to develop tools and some strategic reasoning. One can ask what if I do exercises early in the morning? Will this make me feel better throughout my work routine? There is an intervention here, there is action. The person can try this and learn from the experience. Randomized Controlled Trials are at this rung. One important

detail to keep in mind is that we're talking about the future here, and about a reality that may come true. This one of the main differences between rung 2 and 3. Though in rung 1 we can only see, and in rung 2 we can also do, intervene, in rung 3 we can only imagine. The thing we're imagining will never come true, though in rung 2 what we imagine can happen.

To better understand this, let's look at an experiment that tries to identify if treatment A has a causal effect on some measurement of recovery compared to the usual treatment (let's call it treatment B). Individuals will be randomly assigned either to group 1 (treatment A) or group 2 (treatment B). At the end, we can calculate the Average Treatment Effect (ATE) and, in this hypothetical case we found that there is a causal effect of treatment A in the recovery of patients. They recover more quickly, when compared to treatment B. How would it be to tackle this situation from a rung 3 perspective? We could ask: I see that patient number 1989 received treatment B and took $X$ days to recover. How different would it be if this same patient had received treatment A? There is no possible experiment that would allow us to test this empirically. The patient 1989 received treatment B and the only way to change that would be to travel back in time. Not possible with current technology, as you know. Investigating level 3 questions is much more complicated than level 2 questions and making it clear that these queries are different is therefore very important. The more you go up in the ladder of causation, more assumptions it is required to investigate the event.

## 2.3 Statistical Dependence

When the probability of one event occurring given the occurrence of another event is different from the probability of occurrence of that first event alone, $P(A|B) \neq P(A)$, we say they're dependent, statistically speaking. Usually, it's more common to say that they correlate in some way. Depending on some characteristics of these events, there are many approaches to attempt to estimate numerically this relationship. Even though it has become somewhat common to refer to any sort of relationship as correlation, during this document I will refer to statistical dependence when I'm referring to a statistical relationship in general, and correlation to one of the famous correlation measures, i.e., Pearson correlation coefficient ($\rho$), Spearman's rank correlation ($\rho$) coefficient, or Kendall rank correlation ($\tau$). The important thing to grasp here is that events that have some statistical dependence between them can help us know about one of them when we know something about the other. A wet garden in the morning could help me understand it rained during the night, though my

wife could have woken up earlier and watered it. It's no bullet proof, as you can see. Besides, estimating a statistical dependence between two events, and this can be helpful, don't get me wrong, don't necessarily help me learning how to change one of them.

### 2.3.1 Confounding bias

Hans Reichenbach (1891-1953) introduced the Common Cause Principle and, according to him, if there is some statistical dependence between events $A$ and $B$ in the population, and neither of them cause the other, there must be a third variable $C$ that causes both of them (Reichenbach, 1956). This means that adjusting for this third variable, one will find independence between the two. We know due to the product rule that the joint probability of two events $A$ and $B$ is $P(A, B) = P(A|B)P(B)$ but if $A$ and $B$ are independent, $P(A|B) = P(A)$, that is, knowing about event $B$ does not help me to estimate the probability of event $A$ happening. Based on that, we know that when $A$ and $B$ are independent, their joint probability is $P(A, B) = P(A)P(B)$.

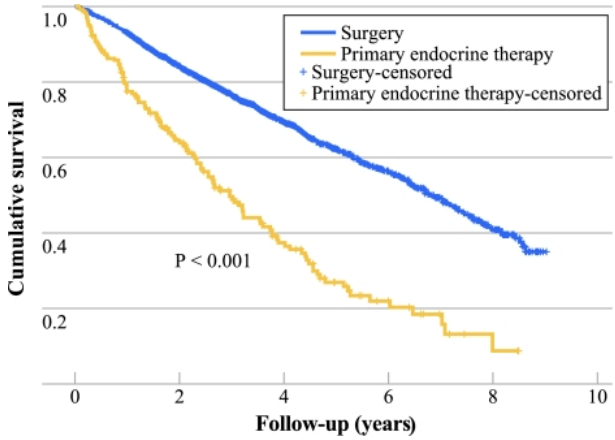
This means that even if we are able to estimate some statistical dependence between $A$ and $B$, and even obtain accurate predictions of $B$ based on measurements of $A$, it does not mean that we can observe a change in $B$ by intervening in $A$, because this relationship could be spurious, non-causal. One famous example mentions shark attacks and icecream consumption, by saying that in some countries in the northern hemisphere it's possible to find a positive correlation between shark attacks, or drownings in beaches, and icecream consumption. The detail here is that both are influenced by season, temperature to be more precise. When it's really warm, people are more inclined to buy icecream and go to the beach, but when it's really cold, people are much less likely to do so. Therefore, when one adjusts by temperature, it's clear the consumption of icecream is not causing the shark attacks or drowning. However, if we ignore temperature, one may feel tempted to create some public policy to stop icecream eating, for example, on the attempt to prevent shark attacks and drownings.

Besides, confounding does not occur in a dichotomous way, either you have dependence or independence. $A$ and $B$ could be causally and directly related and still the accurate estimate be biased due to lack of confounding adjustment. Maybe your treatment seems 10 times better than usual treatment, but it's actually only 2 times better. Or 2 times worse, or has no effect on the disease.

### Confounding by Indication

If we give a quick look at Figure 2.2a, we can see the survival based on what treatment patients aged $\geq$ 75 years old diagnosed with breast cancer followed. At first sight it seems that surgery is better than primary endocrine therapy (PET), which is an oral medication. Clearly patients who are given PET are dying much quicker than those who underwent surgery. You can even think that all patients should undergo surgery and no one should get PET.

Figure 2.2: Overall survival of patients aged $\geq$ 75 years who received primary endocrine treatment in the period 2001–2008 in the south of the Netherlands (n=184) vs. all patients aged $\geq$ 75 years treated with primary surgery in the same region and time period (n=1504)




Retrieved from (Wink et al., 2012)

However, what is not clear in this survival plot is based on what conditions people have been assigned to one treatment over the other. Patients were not given one of the treatments randomly. Table 2.1a shows us that older patients, and with more comorbidities, are the ones receiving PET over surgery, so we're not really talking about similar individuals.

This is what is known as confounding by indication. The outcome we're investigating (survival) is associated with a particular medicine (PET) that is indicated based on other covariates that are confounders. Let's stop thinking about cancer for a second and think of cardiovascular disease (CVD). We're designing an observational study (forget randomization or control group, we're just looking at people in the population) in which we want to see if a particular drug A has an effect on survival of patients. The issue here is that patients with severe cases of CVD are more likely to have a stroke, and therefore more likely to be prescribed drug A. This way, our study may conclude that drug A is bad, but it's actually prescribed for patients

Table 2.1: Age and comorbidity of patients aged $\geq$ 75 years who received surgery or PET for breast cancer in the south of the Netherlands between 2001 and 2008

| Characteristics | Treatment | | | | P value |
|---|---|---|---|---|---|
| | Surgery | (n=1504) | PET | (n=184) | |
| | n | % | n | % | |
| Mean age, y | 80.2 | | 83.8 | | <0.001 |
| No. of comorbidities | | | | | <0.001 |
| 0 | 325 | 21.6 | 16 | 8.3 | |
| 1 | 448 | 29.8 | 51 | 27.7 | |
| $\geq$2 | 574 | 38.2 | 107 | 58.2 | |
| Unknown | 157 | 10.4 | 10 | 5.4 | |

(a) Retrieved from (Wink et al., 2012)

who have severe CVD and that, without drug A, would be in a much worse situation. The lack of effect, of biased effects we observe, are confounding by indication. It's difficult to fight confounding by indication in observational studies because the reasons for drugs being prescribed are usually not recorded in the data (Miettinen, 2011).

### 2.3.2 Collider bias

While confounding bias is related to the common cause principle, a common cause (temperature) distorting the estimate between its effects (icecream consumption and shark attacks), collider bias is related to adjustments to a common effect, a third variable that is caused by $A$ and $B$. The interesting detail is that we usually remove confounding bias by adjusting for the third variable, whereas in collider bias we do not. Actually, the bias occurs when we adjust by this third variable. Being aware of this bias is extremely important because many researchers think that bias can be removed by adjusting for all measured variables. The issue with this practice is that if a measured variable is a collider, by adjusting for it you will be adding bias to your estimate, not removing. Some researchers tend to suggest to not adjust for post-treatment variables, since post-treatment events, that happened after treatment, could not have caused the treatment, that happened before it, and therefore this variable can not be a confounder. Some other authors have invested in making it much more clear when and how we should adjust for variables (Cinelli et al.,

2020b).

One thing that is important to keep in mind is that the adjustment can happen even if the researchers do not do it themselves. Selection bias is one source of collider bias and one famous example of it bias was given by David Sackett (1934-2015), a pioneer in Evidence-Based Medicine (Sackett, 1979). Sackett observed a strong correlation between locomotor disease and respiratory disease in data of 257 hospitalized individuals (odds ratio 4.06). Locomotor disease could lead to inactivity, even sedentarism, which could cause respiratory disease. It could be the case, right? Then he repeated the analysis in a sample of 2783 individuals from the general population and found no association (odds ratio 1.06). It's known that severe locomotor diseases and severe respiratory diseases cause hospitalization, so you have hospitalization as a common effect. If you adjust for this variable, or if you only look at hospitalized patients, you will obtain a spurious estimate of the relationship between the two variables. Sackett called this admission rate bias but it's currently known as selection bias or collider bias. Pearl also referred to this effect as explaining away effect (Kim and Pearl, 1983). It's been shown that about half of experiments in 3 top political science journals introduce bias by conditioning on post-treatment variables, which is something easy to tackle compared to trying to identify if the variable is a collider or not (Montgomery et al., 2018). It's important to remember that just like confounding, the bias produced in this context could not only make independent variables seem dependent, but also distort the dependency. If $A$ and $B$ are causally related, have a common effect $C$ and somehow the researcher is adjusting for $C$, a collider, the statistical dependence between $A$ and $B$ will be biased. Maybe larger than it really is, maybe smaller, maybe zeroed.

### 2.3.3 Correlation measures

There are a few measures of statistical dependence that carry the word correlation with them. The most common ones are Pearson's Correlation Coefficient (PCC), Spearman's rank correlation coefficient and Kendall rank correlation coefficient. They all try in a way or another to capture a type of relationship between the variables being analyzed.

Before talking about correlation, it's important to talk about covariance which is a measure of joint variability between two random variables, based on the idea of variance which is a measure of variability for a single random variable. The covariance of two random variables let us know how these two random variables behave linearly. Apart from the number itself, a positive signal tells us the greater values of one variable correspond to the greater values of the other, while a negative value tells

us the greater values of one correspond to the lesser values of the other. It's tricky to interpret the numbers, the magnitude of the variance, because it depends on the values of the random variables itself, as can be seen in Equation 2.1, for two jointly distributed real-valued random variables $X$ and $Y$ with finite second moments.

$$cov(X, Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))] \tag{2.1}$$

One good idea is to try to normalize the covariance so that it's easier to interpret the value. One way to make it limited to the range $-1$ to $1$ is to divide the covariance by the standard deviation of $X$ and $Y$, and that's the equation for PCC when applied to a population, as can be see in Equation 2.2.

$$\rho_{X,Y} = \frac{cov(X, Y)}{\sigma_x \sigma_y} \tag{2.2}$$

Just like with covariance, PCC investigates the joint variability of random variables that behave linearly. The farther the relationship is to a linear one, the less the interpretation of the coefficient makes sense.

A different correlation measure is Spearman's rank correlation coefficient. It's straightforward to understand it because it is the PCC that we just saw but applied to the rank variables (the ordering of the values of a numeric variable), instead of the variables themselves, as can be seen in Equation 2.3.

$$\rho_{R(X),R(Y)} = \frac{cov(R(X), R(Y))}{\sigma_{R(x)} \sigma_{R(y)}} \tag{2.3}$$

What's interesting about Spearman's correlation is that its interpretation goes beyond linear functions to any sort of monotonic function, which means that some situations in which PCC may not be adequate, Spearman's correlation can be. Besides, even if we're talking about categorical variables, as long as there is an ordering (a discrete ordinal variable), you can calculate the Spearman's correlation.

There are other measures such as Kendal's correlation, Matthew Correlation Coefficient, but they go beyond the scope of this thesis.

#### Partial correlation

So far discussing correlation we have discussed two random variables, but it's true that when talking about confounding and collider bias we mentioned a third variable, and we're all aware that real applications contain many more than two random variables. Partial correlation is what helps us understand the relationship between

some variables taking into consideration a set of other variables.

One common and easy approach to compute the partial Pearson's correlation coefficient is to run linear regressions in order to obtain the residuals and calculate the PCC between the residuals. Let's say we would like to calculate the partial PCC of $X$ and $Y$ given a third variable $Z$. The first step would be to try to fit a line using linear regression to $X$ and $Z$ and obtain, among other things, the residuals, which are the difference between each observed value of the response variable and the value of the response variable predicted from the regression line. One can interpret this as what the linear relationship between $X$ and $Z$ is missing. Do the same for $Y$ and $Z$. Then test if these residuals, $e_x$ and $e_y$, correlate.

### 2.3.4 Information Theory

All the attempts to measure statistical dependence presented so far have limitations regarding the type of relationship between the variables. This subsection introduces a way to measure statistical dependence through Information Theory, more specifically Mutual Information, a measure based on Entropy that does not have such limitations.

#### Entropy

In the context of information theory, the entropy (or Shannon Entropy) of a random variable is the average level of "surprise" related to the possible outcomes of that variable. As an example, consider a fair coin with probability $p = \frac{1}{2}$ of landing on heads and probability $1-p$, which is also $\frac{1}{2}$ of landing on tails. The maximum surprise is for $p = \frac{1}{2}$, when there is no reason to expect one outcome over another. In this case a coin flip has an entropy of one bit. The minimum surprise is for $p = 0$ or $p = 1$, when the event is known and the entropy is zero bits. Other values of $p$ give different entropies between zero and one bits.

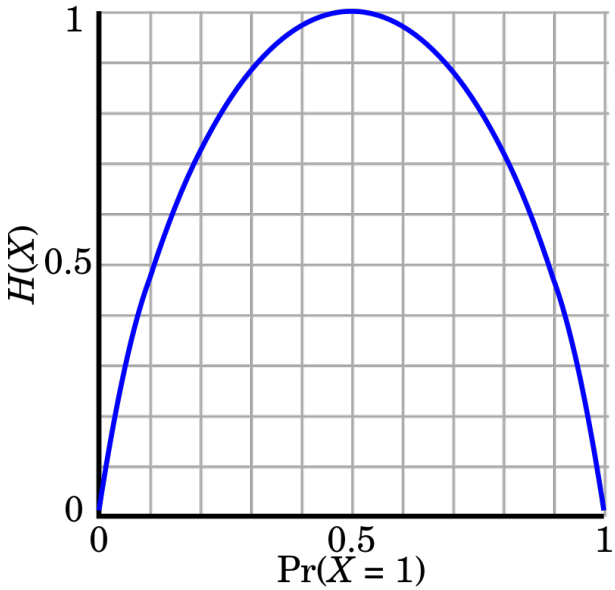
Given a discrete random variable $X$, with possible outcomes $x_1, ..., x_n$, which occur with probabilities $P(x_1), ..., P(x_n)$, the entropy $H$ of $X$ is formally defined as:

$$H(X) = -\sum_{i=1}^{n} P(X_i) log P(X_i) \tag{2.4}$$

The choice of base for log, the logarithm, varies for different applications and base $2$ gives the unit of bits (or "shannons").

As can be seen in Figure 2.3a, for a fair coin Shannon entropy is bounded between $0$ and $1$. The calculation below shows why the maximum uncertainty here is

Figure 2.3: Entropy of a fair coin.




Retrieved from: https://en.wikipedia.org/wiki/File:Binary_entropy_plot.svg

1.

$$H(X) = -\sum_{i=1}^{n} P(X_i) log P(X_i)$$

$$H(X) = -\sum_{i=1}^{2} \frac{1}{2} log_2 \frac{1}{2}$$

$$H(X) = -\sum_{i=1}^{2} \frac{1}{2}(-1)$$

$$H(X) = 1$$

If the coin was biased, which implies $p \neq q$, the maximum entropy will be lower, for the uncertainty is lower than in our previous example. Every time the coin is tossed, one side is more likely than the other, so there is less "surprise". Let's think of $p = 0.7$, for example.

$$H(X) = -\sum_{i=1}^{n} P(X_i) log P(X_i)$$
$$H(X) = -0.7 log_2(0.7) - 0.3 log_2(0.3)$$
$$H(X) = -0.7(-0.515) - 0.3(-1.737)$$
$$H(X) = 0.8816$$

If we think of a lottery, knowing in advance a number that won't be the winning number provides very little information, because most numbers will not be the winning number. But knowing the winning provides a lot of information because it refers to a very low probability event. An equivalent definition of entropy is the expected value of the self-information of a variable. The information content (also called the surprisal) of an event $E$ is a function that increases as the probability $p(E)$ of an event decreases. So when $P(E)$ is close to $1$, we say that the "surprise" of seeing that event is low, but if is close to $0$, we would say that the surprise of seeing that event is high. A possible way to capture that relationship would have been to define the surprise as $\frac{1}{P(E)}$, but in cases when $P(E) = 1$, it would lead to a surprise of $1$ (when it would have made more sense to say it has 0 surprise). Hence, a nicer function to use is the logarithm of 1 over the probability $log(\frac{1}{P(E)})$, which would give us $0$ surprise when the probability of the event is 1.

In fact, it's interesting to know how Shannon developed this equation. He named entropy $H$ (Greek's capital letter eta) after Boltzmann's H-theorem. To understand the meaning of $-\sum p_i log(p_i)$, first define an information function $I$ in terms of an event $i$ with probability $p_i$. The amount of information acquired due to the observation of event $i$ follows from Shannon's solution of the fundamental properties of information:

1. $I(p)$ is monotonically decreasing in $p$: an increase in the probability of an event decreases the information from an observed event, and vice versa.
2. $I(p) \geq 0$: information is a non-negative quantity.
3. $I(p) = 0$: events that always occur do not communicate information.
4. $I(p_1, p_2) = I(p_1) + I(p_2)$: the information learned from independent events is the sum of the information learned from each event.

Given two independent events, if the first event can yield one of $n$ equiprobable outcomes and another has one of $m$ equiprobable outcomes then there are

$mn$ equiprobable outcomes of the joint event. This means that if $log_2(n)$ bits are needed to encode the first value and $log_2(m)$ to encode the second, one needs $log_2(mn) = log_2(m) + log_2(n)$ to encode both. Shannon discovered that a suitable choice of $I$ is given by:

$I(p) = log(\frac{1}{p}) = -log(p)$

In fact, the only possible values of $I$ are $I(u) = k\ log\ u$ for $k < 0$. Additionally, choosing a value for $k$ is equivalent to choosing a value $x > 1$ for $k = -\frac{1}{log\ x}$, so that $x$ corresponds to the base for the logarithm. Thus entropy is characterized by the above four properties.

One interesting thing to notice is that throwing a die has higher entropy than tossing a coin because each outcome of a die toss has smaller probability ($p = \frac{1}{6}$, if it's a six-side fair die) than each outcome of a coin toss ($p = \frac{1}{2}$, if it's a fair coin). Another point is that entropy only takes into account the probability of observing a specific event, so the information that it encapsulates is information about the underlying probability distribution, not the meaning of the events themselves.

#### Differential entropy

Shannon entropy is restricted to a random variable taking discrete values. There are attempts to make it work for a continuous random variable with probability density function $f(x)$ with finite or infinite support $\mathbb{X}$ on the real line by analogy:

$$h(f) = E[-\ln(f(x))] = -\int_{\mathbb{X}} f(x) ln((x)) dx \tag{2.5}$$

This is what is usually called differential entropy (or continuous entropy). Although the analogy between both functions is suggestive, differential entropy lacks a number of properties that the Shannon discrete entropy has. It can even be negative, so is it really Shannon entropy? Corrections have been suggested, such as limiting density of discrete points or using relative entropy.

#### Cross entropy

In the context of information theory, the cross-entropy between two probability distributions $p$ and $q$ over the same underlying set of events measures the average number of bits needed to identify an event drawn from the set if a coding scheme used for the set is optimized for an estimated probability distribution $q$, rather than the true distribution $p$.

The cross-entropy of the distribution $q$ relative to a distribution $p$ over a given set is defined as follows:

$$H(p,q) = -E_p[log\ q] \tag{2.6}$$

where $E_p[\cdot]$ is the expected value operator with respect to the distribution $p$.

The definition may be formulated using the Kullback-Leibler (KL) divergence $D_{KL}(p\ ||\ q)$, divergence of $p$ from $q$ (also known as the relative entropy of $p$ with respect to $q$).

$$H(p,q) = H(p) + D_{KL}(p\ ||\ q) \tag{2.7}$$

where $H(p)$ is the entropy of $p$.

For discrete probability distributions $p$ and $q$ with the same support $\mathbb{X}$ this means:

$$H(p,q) = -\sum_{x \in X} p(x)\ log\ q(x) \tag{2.8}$$

**Joint entropy**

The notation $H(p,q)$ is also used for a different concept, the joint entropy of $p$ and $q$ and in many situations this can be misleading for the two concepts mean different things. Just like entropy is a measure of uncertainty associated with a variable, joint entropy is a measure of uncertainty associated with a set of variables.

The joint Shannon entropy (in bits) of two discrete random variables $X$ and $Y$ with images $\mathcal{X}$ and $\mathcal{Y}$ is defined as:

$$H(X,Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(x,y) log_2[P(x,y)] \tag{2.9}$$

One can also deal with more than two variables. The expansion then is:

$$H(X_1,...,X_n) = -\sum_{x_1 \in \mathcal{X}_\infty} \sum_{x_n \in \mathcal{X}_\backslash} P(x_1,...,x_n) log_2[P(x_1,...,x_n)] \tag{2.10}$$

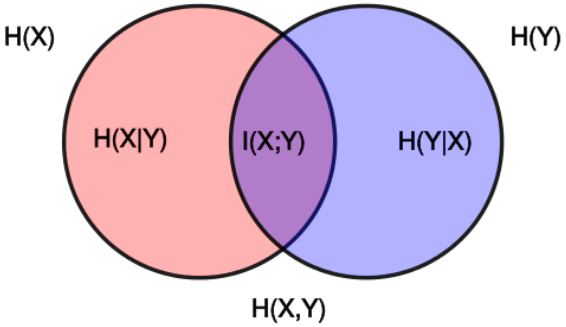
Some interesting properties of the joint entropy include nonnegativity ($H(X_1,...,X_n) \geq 0$), and it is always greater than or equal to any of the individual entropies ($H(X_1,...,X_n) \geq \max_{1 \leq i \leq n} H(X_i)$).

### Mutual Information

We can define the mutual information $I$ of $X$ and $Y$, in terms of Shannon entropy, as:

$$I(X;Y) = H(X) - H(X|Y) \tag{2.11}$$

Figure 2.4: Venn Diagram.




Retrieved from: https://en.wikipedia.org/wiki/File:Entropy-mutual-information-relative-entropy-relation-diagram.svg

The Venn Diagram in Figure 2.4a is helpful to understand Shannon Entropy and Mutual Information. The whole circle on the left, $H(X)$, minus the red-only part, $H(X|Y)$, is the mutual information between $X$ and $Y$, e.g., $I(X;Y)$, as seen in Equation 2.11. The $H(X,Y)$ below the middle is a bit confusing, since it is not referring to the middle part but for the join entropy of $X$ and $Y$, that is, both circles. And because $H(X,Y) = H(X) + H(Y) - I(X;Y)$ (so that the middle part is not counted twice) you can also find the mutual information between $X$ and $Y$ through the following equation:

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{2.12}$$

Though it's nice to understand mutual information in terms of Shannon entropy and with a Venn diagram you can also calculate it in terms of what Shannon entropy really makes use of, that is, the probability of levels in random variables. Besides, one can also make use of KL divergence to test the difference between two distributions. Let $(X,Y)$ be a pair of random variables, with values over the space $\mathcal{X} \times \mathcal{Y}$, with joint distribution $P_{(X,Y)}$ and marginal distributions $P_X$ and $P_Y$. Then:

$$I(X;Y) = D_{KL}(P_{(X,Y)}||P_X P_Y) \tag{2.13}$$

It's intuitive to see why this works. If $P(X,Y) = P(X)P(Y)$, it's because $X$ and $Y$ are independent. Mutual Information has the following properties:

- Nonnegativity: $I(X;Y) \geq 0$
- Symmetry: $I(X;Y) = I(Y;X)$
- Relation to marginal, conditional and joint entropy:

$$\begin{aligned} I(X;Y) &\equiv H(X) - H(X|Y) \\ &\equiv H(Y) - H(Y|X) \\ &\equiv H(X) + H(Y) - H(X,Y) \\ &\equiv H(X,Y) - H(X|Y) - H(Y|X) \end{aligned}$$

As it's been shown so far, Mutual Information is not limited to a specific type of relationship between the random variables. It really measures statistical dependence.

#### Conditional Mutual Information

In some circumstances, it may be useful to express the mutual information of two random variables conditioned on a set of other variables. In terms of marginal and joint Shannon Entropies, the mutual information of $X$ and $Y$ given a third variable $Z$ can be expressed as:

$$I(X;Y|Z) = H(X,Z) + H(Y,Z) - H(X,Y,Z) - H(Z) \tag{2.14}$$

And also as:

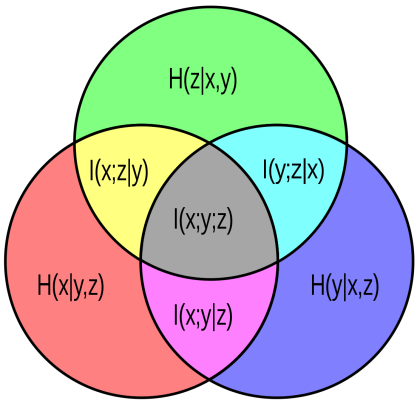
$$\begin{aligned} I(X;Y|Z) &\equiv H(X|Z) - H(X|Y,Z) \\ &\equiv H(X|Z) + H(Y|Z) - H(X,Y|Z) \end{aligned}$$

And even with only Mutual Information expressions:

$$\begin{aligned} I(X;Y|Z) &\equiv I(X;\{Y,Z\}) - I(X;Z) \\ &\equiv I(Y;\{X,Z\}) - I(Y;Z) \end{aligned}$$

The second equation in the block above is more commonly seen, called *the chain rule for Mutual Information*. I bring again another Venn diagram in Figure 2.5a that can help us understand what's going on with these expressions.

Figure 2.5: Venn Diagram.




Retrieved from: https://en.wikipedia.org/wiki/Conditional_mutual_information#/media/File:VennInfo3Var.svg

The whole circle on the left is $H(X)$, the whole circle on the right is $H(Y)$ and the whole circle above is $H(Z)$. The intersection pink + grey is $I(X;Y)$. The intersection grey + light blue is $I(Y;Z)$ and the intersection grey + yellow is $I(X;Z)$. Based on this, we know that if we subtract $I(X;Y|Z)$ from $I(X;Y)$ we will get $I(X;Y;Z)$ and indeed this is a commonly used equation for $I(X;Y;Z)$.

$$I(X;Y;Z) = I(X;Y) - I(X;Y|Z) \tag{2.15}$$

The $I(X;Y;Z)$ is known as 3-point Mutual Information or Interaction information, as just like Mutual Information, it is symmetric so it does not matter which variable is conditioned on. It's also bounded, according to the equation below:

$$-min(I(X;Y|Z), I(Y;Z|X), I(X;Z|Y)) \leq I(X;Y;Z) \leq min(I(X;Y), I(Y;Z), I(X;Z)) \tag{2.16}$$

One interesting thing here is that, differently from Mutual Information and Entropy, 3-point Mutual Information can be negative! Keep this in mind, as this result will be very useful when we talk about causal discovery and v-structures.

## 2.4 Structural and Graphical Causal Model

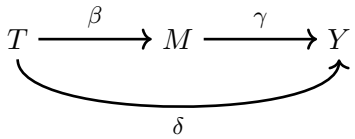
A causal model, short for structural causal model (SCM), is an ordered triple $\langle U, V, E \rangle$ where $U$ is the set of exogenous variables (variables caused by events outside the model), $V$ is the set of endogenous variables (variables caused by events contained

in the model), and $E$ is the set of structural equations that relate the variables in $U$ and $V$. These are called structural equations, and not simply equations, because they're not symmetric. They're specifically referring to the data generating process, how each variable was created, that is which variables caused it. The goal of causal models is to investigate the causal relationship between entities, though it can also be useful to improve study designs and other types of analyses by providing insight into what is the role of certain variables, how they should be adjusted or experimented upon.

A similar thing to a SCM is a Graphical Causal Model in which we have graphs representing the functional relationships between the variables. The next subsection presents the required knowledge of graph theory to understand the applications of causal models that will be seen in next sections of this work.

### 2.4.1 Graph Theory

A graph is a mathematical structure $G(V, E)$ where $V$ is the non-empty set of objects, called vertices, and $E$ is a set of unordered pairs of vertices in $V$, representing the relationship (or lack thereof) between vertices, also known as nodes. If $E$ is a set of ordered pairs, we have a directed edge and the dash (—) will become an arrow ($\rightarrow$). When we have orientation, one can classify nodes as parents of a node $X$ in $V$ if the nodes are oriented towards $X$ ($Y \rightarrow X$ and $Z \rightarrow X$) or descendants of $X$ if $X$ is oriented towards them ($X \rightarrow Y$ and $X \rightarrow Z$). Below, you can see one example of a directed graph with nodes $T$, $M$ and $Y$ and edges from $T$ to $M$, from $T$ to $Y$ and from $M$ to $Y$. If this is a linear model, the $\beta$, $\gamma$, and $\delta$ are the linear coefficients of the equations.




The graph above is equivalent to the following structural equations, where $U_1$ is an unobserved cause of $T$, or exogenous variable and the structural equations for our endogenous variables $M$ and $Y$ explain what causes them:

$$T = U_1$$
$$M = \beta T$$
$$Y = \delta T + \gamma M$$

A cycle happens when following the path through directed edges you can arrive at a vertex that has already been reached through the path. If the path is simply connecting one edge to itself, this cycle can also be called a loop. In causal modeling, it's very common to make use of a class of graphs in which all edges are oriented and there are no cycles. A graph in this class is known as a Directed Acyclic Graph (DAG). However, some algorithms and analysis work with a different class of graphs called Completed Partially Directed Acyclic Graph (CPDAG). In CPDAGs, differently from DAGs, there is not only one type of edge (which is an edge with an orientation, from left to right, $\rightarrow$, for example), but also an undirected edge ($-$) illustrating ambiguity (it could be $\leftarrow$ or $\rightarrow$). That's the meaning of the *partially oriented* graph. CPDAGs are useful because they provide an equivalence class of graphs, differently from DAGs that provide one unique graph. The CPDAG with two nodes $A$ and $B$, $A - B$, is an equivalence class that includes two DAGs: $A \leftarrow B$ and $A \rightarrow B$.

There is a third class of graphs called Maximal Ancestry Graph (MAG). In MAGs, we have oriented graphs, like in DAGs, undirected edges like in CPDAGs but also bi-directed edges, $A \longleftrightarrow B$, that are equivalent to the presence of an unobserved (latent) common cause of the two connected variables, $A \leftarrow\!\!- L \dashrightarrow B$. However, just like for DAGs, it's not easy to find a unique MAG, and therefore there is an equivalence class of graphs called PAG (Partial Ancestry Graph). In this class, there are 6 kinds of edges (pay attention to the two extremities of the edge) with three symbols:

- Circle: In the PAG equivalence class, there is at least one MAG in which this extremity of the edge is the tail (such as the left in $\rightarrow$) and at least one in which this extremity is an arrowhead ($\leftarrow$).
- Blank (tail): In the PAG equivalence class, all MAGs have this extremity of the edge as a tail (such as the left side of $\rightarrow$).
- Arrowhead: In the PAG equivalence class, all MAGs have this extremity of the edge as an arrowhead.

The six edges are then: circle-circle, circle-blank, circle-arrow, blank-arrow, arrow-arrow, and blank-blank. In Section 2.5, I will talk about algorithms that learn these structures from data, but it's useful to mention some of them as examples as for which classes of graphs they provide as output of their computation. The PC and GES algorithm (Chickering, 2002, Spirtes et al., 2000), for example, output a CPDAG, while the FCI and RFCI algorithms output a PAG (Colombo et al., 2012, Spirtes et al., 1999, 2000), when the underlying graph is a DAG or a MAG, respectively, and that the causal discovery is fully correct (which requires an infinite sample size in general).

## 2.5 Network Inference and Causal Discovery

The task of learning the structure of a graph from data is called network inference. When we're not interested in mere statistical relationships, but causal relationships, this is called causal discovery or exploratory causal analysis. In many fields, such as epidemiology, econometrics and artificial intelligence, it's common to use graphs to investigate causality (Cunningham, 2021, Hernán and Robins, 2020, Pearl, 2000). It's useful to visually understand the relationship between variables and Pearl and others developed many tools to investigate causal relationships through graphs, such as the backdoor criterion, frontdoor criterion, among other for causal identification. If the causal effect that we want to estimate does not pass such criteria, it's not identifiable and we have red light regarding advancing for the next stage, which is estimation.

It's common for many people to be excited when they learn about causal graphs, how easy it is to visualize the relationships, test for identifiability and decide what are good and bad controls (Cinelli et al., 2020b). However, right after that, a doubt usually follows: But how does one get these graphs? It's arbitrary? Obviously it must not be arbitrary. Sometimes, background knowledge of domain experts along with experts in graph theory get together to draw a graph that makes sense to them and then they test how compatible the data are with the graph. In other situations, the Causal Discovery approach, or Exploratory Causal Analysis (ECA) is used to learn the causal structure through data. When the data are non-experimental, or observational, this is called causal discovery with observational data.

Causal Discovery and network inference in general is an example of an inverse problem. Such problems are very interesting because they help us estimating parameters that we could not directly observe. At the same time, inverse problems are very challenging as many different things could have led to what we observed. In the case of Causal Discovery, in many circumstances there is more than one graph that could fit the data. Assumptions can be used to shorten the set of possible graphs. In the following subsections, many methods for network inference will be presented, starting with simple ideas and advancing to more complex ones and famous algorithms.

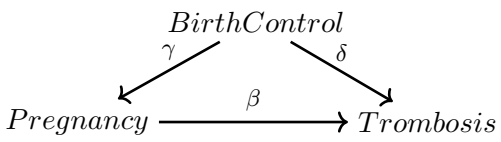
### 2.5.1 Assumptions for Causal Discovery

There are three main assumptions that are crucial to allow us to perform network inference. In the context of causal discovery, the first one is usually referred to as Causal Markov Condition or Causal Markov Assumption. If a graph $G$ satisfies the

Markov Condition, we know three things:

- Pairwise Markov property: Any two non-adjacent nodes are conditionally independent given all other nodes. $X_u \perp\!\!\!\perp X_v | X_{V/_{\{u,v\}}}$
- Local Markov property: A variable is conditionally independent of all other variables given its neighbors (parents and descendants). $X_v \perp\!\!\!\perp X_{V/N[v]} | X_{N(v)}$;
- Global Markov property: Any two subsets of variables are conditionally. independent given a separating subset. $X_A \perp\!\!\!\perp X_B | X_S$

In the context of causality, the local Markov property is known as Causal Markov condition and if the graph $G$ is a DAG, this condition is equivalent to the global Markov property (which allows us to make nodes independent given a separation set). The Causal Markov condition is usually expressed in terms of node $A$ being independent of everything but its descendants (effects), given its parents (causes). It's the Markov condition that allows us to go from statistical distributions to edges in a graph. However, this assumption can be violated due to unmeasured confounders, that is, we may be inclined to believe that there is dependence between $A$ and $B$, and therefore an edge should connect them, when actually there is a third unmeasured variable that causes $A$ and $B$, it's a parent of both of them, an effect known as confounding. If we had measured this third variable $C$, we could make $A$ and $B$ independent given $C$, but we didn't so the true graph does not satisfy the Markov condition. A second assumption is called Causal Sufficiency that states that we have measured all common causes of the measured variables. Assuming causal sufficiency, we can assume the causal Markov condition. The third and last main assumption is Causal Faithfulness. One can see Faithfulness as the opposite of the Markov condition. The Markov condition allows us to put an edge when there is statistical dependence, and faithfulness allows us to know there is statistical dependence when we see an edge. In other words, there are no independencies between variables that are not represented by the Causal Markov Condition. This may seem obvious, but this assumption can also be violated! In the graph below, considering it's a linear model, if $\gamma\beta = \delta$, the direct effect from $BirthControl$ to $Trombosis$ ($\delta$) will be cancelled by the indirect effect, mediated through $Pregnancy$ ($\gamma\beta$).



It's usually assumed that these perfect cancellations are rare and that thus we should not worry about causal faithfulness, but in some self regulatory systems (homeostasis) in biological organisms, it's been shown that it's less rare than initially expected, so it's something to keep in mind. Now that graph theory has been introduced to some degree, and also the assumptions for causal discovery, there are three famous graphoids that must be introduced, mostly because they graphically illustrate what we've seen in previous sections through words and mathematical expressions. They're unshielded triplets, which means that we have three edges and only two edges.

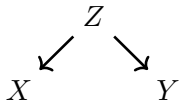



Figure 2.6: This structure is known as fork. $Z$ is a confounder. If you don't consider/measure $Z$, you will observe some statistical dependence between $X$ and $Y$ but there will be none once you take $Z$ into consideration, that is, $X \not\perp\!\!\!\perp Y$, but $X \perp\!\!\!\perp Y|Z$.

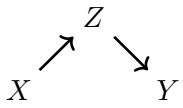



Figure 2.7: This structure is known as chain. $Z$ is a mediator. If you don't consider/measure $Z$, you will observe some statistical dependence between $X$ and $Y$ but there will be none once you take $Z$ into consideration, that is, $X \not\perp\!\!\!\perp Y$, but $X \perp\!\!\!\perp Y|Z$. It's important to notice that the same thing would have happened if we had instead $X \leftarrow Z \leftarrow Y$.

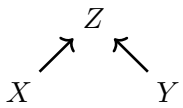



Figure 2.8: This structure is known as V-structure. $Z$ is a collider. If you don't consider/measure $Z$, you will observe no statistical dependence between $X$ and $Y$ but there will be some dependence once you take $Z$ into consideration, that is, $X \perp\!\!\!\perp Y$, but $X \not\perp\!\!\!\perp Y|Z$.

Two important remarks can be made based on what we just saw in Figures 2.6, 2.7 and 2.8:

1. If we have three variables $X$, $Y$, and $Z$, we observe some statistical dependence between $X$ and $Y$ but none when we adjust for $Z$, we can't know if the

graph structure is a fork or chain. We can only get to an equivalence class that contains three graphs: the fork and two equivalent chains.

2. But if we have three variables $X, Y$ and $Z$, with $X \perp\!\!\!\perp Y$ but $X \not\perp\!\!\!\perp Y|Z$, then there is only one graph that could fit such relationships taken into consideration the causal Markov condition and Causal Sufficiency: The v-structure, also known as signature of causality. Many algorithms take advantage of this, as we will soon see.

### 2.5.2 Simple approaches for network inference

#### Covariance Matrix

Covariance has already been mentioned earlier when correlation measures were presented, in subsection 2.3.3 (Equation 2.1). The covariance matrix, also known as auto-covariance matrix or dispersion matrix, is a square matrix in which every cell contains the covariance between the random variable in row $i$ and the random variable in column $j$. When the random variable in row $i$ and column $j$ are the same, we have $cov(X, X) = var(X)$ so the main diagonal of the covariance matrix is the variance of the random variables, and that's why this matrix is also known as variance-covariance matrix. Besides, the covariance matrix will always be symmetric and positive semi-definite ($x_{i,j} \geq 0, \forall i, j$). The covariance matrix of a random vector $X$ is typically referred as $K_{XX}$.

If the entries in the column vector $X = (X_1, X_2, ..., X_n)^T$ are random variables, each with finite variance and expected value, the covariance matrix $K_{XX}$ is the matrix whose $(i, j)$ entry is the covariance and can be calculated as:

$$K_{X_iY_j} = cov[X_i, X_j] = E[(X_i - E[X_i])(X_j - E[X_j])] \tag{2.17}$$

In the equation above the operator $E$ denotes the expected value (mean) of its argument. The auto-covariance can be expressed as:

$$K_{XX} = cov[X, X] = E[(X - \mu x)(X - \mu x)^T] = E[XX^T] - \mu x \mu x^T \tag{2.18}$$

where $\mu x = E[X]$. The auto-covariance matrix $K_{XX}$ is related to the autocorrelation matrix $R_{XX}$ by $K_{XX} = E[(X - E[X])(X - E[X]^T] = R_{XX} - E[X]E[X]^T$ where the autocorrelation matrix is defined as $R_{XX} = E[XX^T]$. The inverse of this matrix, $K_{XX}^{-1}$, if it exists, is the inverse covariance matrix, also known as the concentration matrix or precision matrix. The precision matrix is usually much sparser, when com-

pared to the covariance matrix, because it tends to have $0$ for cases in which the two random variables are conditionally independent. Some authors prefer to work with the precision matrix, instead of the covariance matrix, due to some other interesting properties (Bernardo and Smith, 2009). For example, in Bayesian statistics, if both the prior and the likelihood have Gaussian form, and the precision matrix of both of these exist (because their covariance matrix is full rank and thus invertible), then the precision matrix of the posterior will simply be the sum of the precision matrices of the prior and the likelihood.

The idea is that by having matrices of linear relationships, we can identify which random variables have a relationship with another and based on that we can construct a graph. Random variables will be nodes in the network and nodes $i$ and $j$ that have a value $0$ in the cell $i, j$ of the matrix will have no edge between them. Otherwise, an edge will be there. Some authors decide on a threshold and values below this threshold will mean no edge. This is required as getting an exact value of $0$ is difficult in practice.

As mentioned before, if a cell value of $0$ in the precision matrix, or even the covariance matrix, means a lack of linear relationship between these two variables, and in the case of precision matrix this is more common because it refers to a lack of linear relationship conditioning on other variables, it's clear that the cell values for the covariance matrix are not restricted to the direct relationship between the two random variables alone. There's bias! In a similar situation before, the concept of partial correlation was mentioned and again that's what we'll show here. If we want the covariance matrix $K_{X,Y}$ (the covariance between two vectors of random variables $X$ and $Y$) to not have the relationships considering a possible relationship with a third vector of random variables $A$, we will calculate the partial covariance matrix $K_{X,X|A}$ with the equation below:

$$K_{X,Y|A} = pcov(X, Y|A) = cov(X, Y) - cov(X, A)cov(A, A)^{-1}cov(A, Y) \tag{2.19}$$

In the context of causality, the main limitation seen with this approach is that it usually does not differentiate between good and bad controls and, as we've already seen, adjusting for colliders will add bias to our analyses, instead of removing bias. Besides, it's limited to linear relationships. With that in mind, other approaches have tried to make use of statistical independence (and conditional independence) to decide if two variables are related or not. The idea follows that $X \perp\!\!\!\perp Y \longleftrightarrow P(X, Y) = P(X)P(Y)$, and for the case of conditional independence, $X \perp\!\!\!\perp Y|Z \longleftrightarrow P(X, Y|Z) = P(X|Z)P(Y|Z)$.

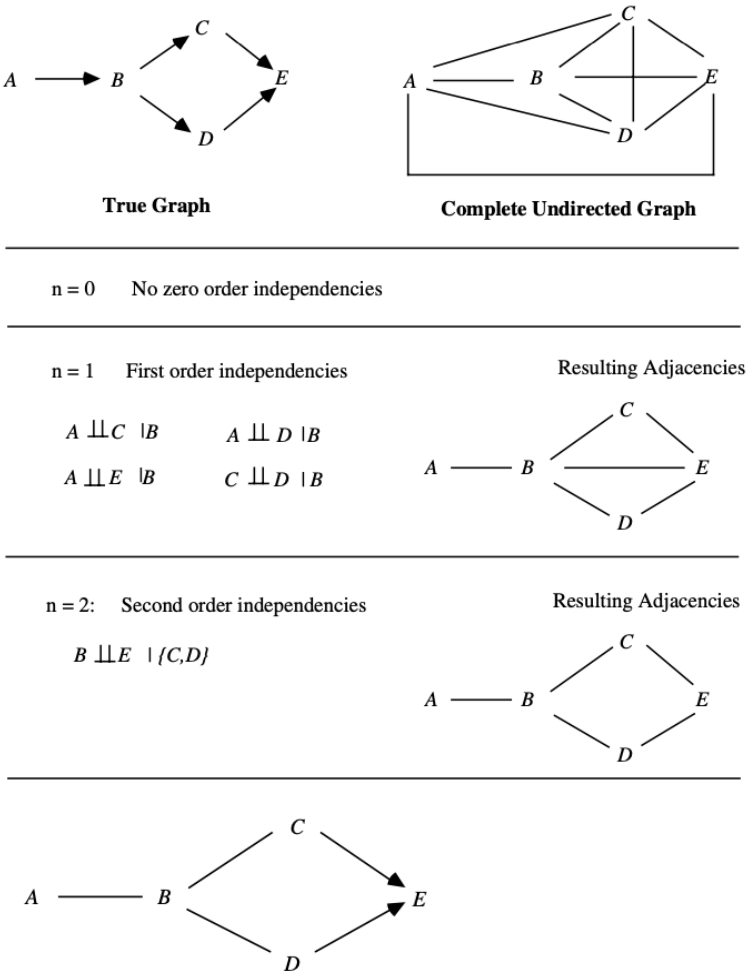
### 2.5.3 Constraint-based methods

Most constraint-based methods for network inference make use of independence tests. This means that as long as a non-parametric independence test is used, the network inference approach will also be non-parametric. These methods usually rely on the causal Markov condition, causal sufficiency and causal faithfulness. The most famous constraint-based method for causal discovery is the PC algorithm and its derivations that will be shortly described below.

#### The PC algorithm

There is a famous algorithm called PC that makes use of this idea (Spirtes et al., 2000) and is illustrated in Figure 2.9a. The true graph is shown and we start with a complete undirected graph (all nodes are connected to all nodes). The first step is to remove all edges between nodes that are marginally independent (the set of variables to condition on is empty, i.e. $P(X, Y) = P(X)P(Y)$). For the example illustrated in Figure 2.9a, there is none. The next step is to try to make nodes independent by adjusting on neighboring nodes. The edge between $A$ and $C$ is removed because $P(A, C|B) = P(A|B)P(C|B)$, that is $A$ and $C$ are independent given $B$. The same thing happens for the edge between $A$ and $D$, $A$ and $E$, and $C$ and $D$. All these tests were done with only one conditioning variable. The next step is to go to two variables in the adjustment set. $B$ and $E$ are independent given $C$ and $D$. And this keeps going. When this is over, we have the skeleton and no edges will be added or removed after this time. At this point, the orientation part starts. The algorithm will look for v-structures and orient them as such, and then will orient the remaining edges according to two rules: No new v-structures and no cycles (after all, the final goal of PC is to infer a Directed Acyclic Graph). PC is one of the most famous constraint-based algorithms.

Figure 2.9: PC Algorithm




Retrieved from: (Spirtes et al., 2000)

One of the most famous limitations of PC is that noisy/finite data could lead the independence test to give a different result, when compared to the true graph, and such early mistakes propagate to other tests and decisions during the rest of the algorithm. The lexicographical order in which the variables are tested for independence would lead to different results. A later version of PC called PC-Stable made it order-independent (Colombo et al., 2014). Changes to the orientation step have also been made, known as conservative rule and marjority rule (Colombo et al., 2014, Ramsey et al., 2012). With the conservative rule, V-structures are only oriented when the collider $Z$ is in neither of the separation sets that satisfy $X \perp\!\!\!\perp Y|U_i$. With the majority rule, as long as the collider $Z$ is in $U_i$ less than $50\%$ of time, it's OK to orient the v-structure.

Another limitation of the PC algorithm is that the separation sets in the final network sometimes were not consistent to the definition of d-separation in graphs (Li et al., 2019). For example, $X$ and $Y$ were made independent given $Z$ but $Z$ is not a mediator or a confounder of $X$ and $Y$, and therefore should be in the separation set.

#### MXM

MXM is a constraint-based method for handling mixed variables (Tsagris et al., 2018). It can work as a plugin to other famous constraint-based methods such as PC or FCI, allowing them to reconstruct causal networks with mixed variables. It makes use of likelihood-ratio tests based on regression models to estimate conditional independence for variables. The likelihood-ratio test for conditional independence between $X$ and $Y$ given a set $Z$ can be performed by fitting regressions in different ways and comparing their goodness-of-fit. If they're really independent, the models should fit equally well as the inclusion of extra variables shouldn't provide any additional information for the relationship between $X$ and $Y$.

### 2.5.4 Score-based methods

Differently from constraint-based methods, score-based methods try to fit graphs from an equivalence class to the available data and score this fit. In the end, the graph with the best fit, using criterion such as BIC (Bayesian Information Criterion) or AIC (Akaike Information Criterio), for example, is chosen. It's clear to see how tricky this can be as the search space grows exponentially with the number of nodes. Some approaches make use of greedy algorithms or some heuristic to limit the search space and provide an adequate solution in reasonable time.

Given the data $\mathcal{D}$ from a vector of variables $V$, find the graph $\widetilde{\mathcal{G}}$ that maximizes a likelihood score $S(\mathcal{D}, \widetilde{\mathcal{G}})$ for each $\mathcal{G}$ in the space of DAGs:

$$\widetilde{\mathcal{G}} = argmax_{\mathcal{G}} S(\mathcal{D}, \mathcal{G}) \tag{2.20}$$

#### GES

The Greedy Equivalence Search (GES) algorithm is one of the most famous score-based methods for causal discovery (Chickering, 2002). Differently from PC that starts with a complete unoriented graph, a skeleton, GES starts with an empty graph, no edges, just the nodes. At every iteration, there is a forward phase in which it's

tested if it's possible to increase the score by adding an edge between two nodes. If it's possible, do it. Then, there is the backward phase, in which it's tested if it's possible to increase the score by removing an edge between any two nodes. If it's possible, do it. When the score stops increasing, the algorithm stops.

### 2.5.5 Other approaches

#### LiNGAM

Shimizu et al. (2006) developed an approach to recover the causal structure in a very different way, by noticing that for non-Gaussian distributions, there is more information in the joint distribution than in the covariance matrix alone, and that this can be detected making use of Independent Component Analysis (ICA). LiNGAM stands for Linear non-Gaussian and Acylic Model and the pairwise relationship is modeled through a linear structural equation, such as:

$$Y = bX + \epsilon \tag{2.21}$$

with $\epsilon$ being some noise and $\epsilon \perp\!\!\!\perp X$. Whenever at most one $\epsilon$ (the exogenous variables) is Gaussian, it was shown that it is possible to recover the structure, as long as the other assumptions are not violated, such as linearity (Shimizu et al., 2006). An improvement known as DirectLiNGAM brought a new way to identify the causal direction between nodes through the use of regressions and independence tests between the predictors and residuals (Shimizu et al., 2011).

#### CausalMGM

CausalMGM (Causal Mixed Graphical Models) is a hybrid approach, making use of both constraint-based and score-based ideas (Sedgewick et al., 2016, 2019). It works with both discrete and continuous variables, and seeks to output a non-directed graphical model in which edges represent variables that are not conditionally independent given the other variables in the graph. An edge, therefore, does not mean that $X$ and $Y$ are causally related in a direct way. It can mean that $X$ causes $Y$, $Y$ causes $X$, $Z$ causes $X$ and $Y$ with $X \perp\!\!\!\perp Y$, or selection bias (collider adjustment) making $X$ and $Y$ dependent when they're actually independent. It provides a web application with preprocessing features, non-directed graphical reconstruction with MGM and a third optional step to identify the orientations of the edges based on PC-Stable. It has many limitations such as not supporting missing data, interactions between continuous variables are assumed to be linear, and these variables

must be close to normally distributed. They show that imposing separate sparsity penalties for edges connecting different types of nodes significantly improves edge recovery performance. For model selection, instead of using Akaike Information Criterion or Bayesian Information Criterion, they developed a method called Stable Edge-specific Penalty Selection.

## 2.6 Paradoxes and fallacies related to causality

There are many examples that for a long time had been seen as paradoxical but became trivial to understand when viewed through the lens of causality and the use of causal graphs. This section highlights a few, while referring to concepts previously presented such as confounding, in subsection 2.3.1, and collider bias in subsection 2.3.2.

### 2.6.1 Berkson's Paradox

Berkson's Paradox is actually a veridical paradox, because even though it's counter intuitive, it's actually accurate. It happens due to collider adjustment and in many situations the implicit adjustment leads the analyst to believe the result to be paradoxical or incorrect. Three famous examples are the Low-birth weight paradox, when the adjustment is explicit, the Monty Hall problem when it's explicit but difficult to identify and the study conducted by Sackett mentioned in subsection 2.3.2, which was implicit (Sackett, 1979).

#### Low-birth weight paradox

The low-birth weight paradox was initially described in the 60's by Jacob Yerushalmy (1904-1973), and later published as a scientific article (Yerushalmy, 1971). It's surprising to realize that only in 2006 it was possible to fully understand and explain his paradoxical observations (Hernández-Díaz et al., 2006). For a very long time, it's been common practice to record birth weight. Not only because it is easy to measure, but because there is a correlation between infant mortality and birth weight. It's been agreed that birth weight below $2.5$kg is considered low birth weight and indeed infants that are born with $< 2.5$kg have a much higher chance of dying in the first weeks than those that are born with $\geq 2.5$kg. Besides, infants whose mothers smoked during pregnancy are more likely to be born with low birth weight and it made a lot of sense that these infants would have worse mortality than other infants whose mothers did not smoke. Actually, a nationwide study in the US at the

time showed that babies that were born with less than $2.5$kg had a death rate over 20 times worse than that of infants that were born with non-low birth weight.

Yerushalmy, a biostatistician, conducted starting in 1959 a long-term public health study that collected pre- and postnatal data on over 15,000 children in the San Francisco Bay Area. His data also showed that infants from mothers that smoked during pregnancy were lighter on average than the babies of non-smoking mothers. However, to his surprise, low birth weight infants of smoking mothers had a better survival rate than low birth weight infants form non-smoking mothers. His analysis was suggesting that smoking had a protective effect! Here, however, Yerushalmy was consciously and explicitly adjusting for a collider as can be seen in Figure 2.10. He was only looking at low birth weight infants (this is also known as selection bias). The dashed rectangular box identifies adjustment.

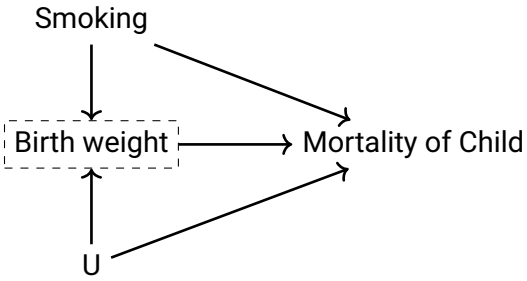



Figure 2.10: Smoking during pregnancy has a direct causal effect on mortality of child and an indirect effect through the effect on birth weight, i.e. Smoking $\rightarrow$ Mortality of Child, and Smoking $\rightarrow$ Birth weight $\rightarrow$ Mortality of Child. On the other hand, smoking during pregnancy is not the only cause for low birth weight. There are more severe causes for this that not only have a direct effect on mortality but also an indirect one through birth weight. These unmeasured other causes are named U here and one example is Congenital Brain Injury. It's possible to see that birth weight is a collider and if you only look at babies with low birth weight, you're conditioning on this node and therefore adding spurious dependence to your estimate.

This is another example of the explain away effect discussed in subsection 2.3.2. If the infant has low birth weight and the mother smoked during pregnancy, there is a smaller chance that other causes for low birth weight occurred, and vice-versa. The explanation here is not that smoking has a protective effect, but that for low birth weight there are more severe causes. Looking at all babies, with low and non low birth weight, it's possible to see that smoking leads to higher infant mortality. The low birth weight paradox is an example of how bad variable adjustment can lead us to spend decades lost in discussions about something that looks paradoxical but in the end is not.

**Monty Hall Problem**

There was a TV show in American Television called *Let's Make a Deal*. In it, there was a game that presented three doors to participants and behind one of them there was a nice prize. In the other two, though, there was something worthless. In many times, the participant would choose a door, Monty Hall would ask if the person was sure, before opening it, and if the person was certain about the picked door, he would open it. However, in some other times, Monty Hall would open a second door, before opening the first picked by the participant, showing a worthless prize (if he picked the door for the nice prize, the game would be over) and then ask if the person would still want to stick with the first door chosen. Most people at the time, including statisticians and mathematicians, believed the chance to find at random the nice prize behind any of the three doors was 1/3 (Mlodinow, 2009). This wouldn't change if a second door was opened, but they were wrong! Computer simulations later showed that what some people suggested was right: You increase your chance of winning by changing doors. You will not always win, of course, but changing doors makes you more likely to win. Figure 2.11 shows the DAG for the Monty Hall game.

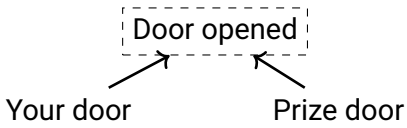



Figure 2.11: There are three doors. In this diagram, it's shown the door you chose, the door Monty Hall opens and the door in which the prize is located. It makes no sense for Monty Hall to open the door you chose, or the door where the prize is hidden. So there is an arrow from your first door and the prize door directing him to the only door he can open. This means, that even though the door you chose and the door with the prize are independent, when you know which door he opened (you adjust for a collider), there is some spurious correlation between your door and the prize door and that's why you have a higher chance of changing.

Why doesn't this sound intuitive? Why it's so hard to grasp what's going on? This happens mostly because in our mind, we think of a different game. We think about the game in Figure 2.12, in which Monty Hall doesn't know where the prize is, so knowing the door he opened doesn't help us with anything. Door opened is no long a collider.

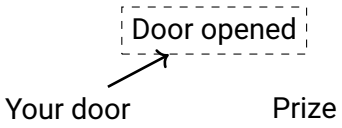



Figure 2.12: monty hall

Before proceeding, I want to make sure that the explain away effect is really clear. Let's think of a simpler example in which there are only two possible ways for a car not to start: Either the car battery is dead or there is no fuel. In other words, there are two events that can lead the car not to start. See the Figure 2.13.

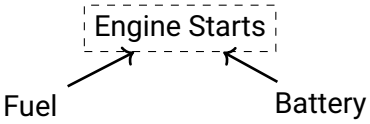



Figure 2.13: Even if you understand that lack of fuel or a discharged battery can lead a car engine not to start (only one of them, or both together), knowing information about one of them doesn't help you predict the other. If you check the fuel tank of a car and I ask you to predict if the battery is charged or not, there is no information to help you. You're just clueless. Why? Because fuel tank and battery being charged are independent. But if you checked the fuel tank and there is fuel, and I tried to turn on the car and it didn't start, you now know that the battery is discharged because it's the only possibility. Knowing one, and adjusting for a collider (the car doesn't start) explains away the possibility of being lack of fuel: It has to be a discharged battery!

#### Sackett's study

Now we can see with graphical models what we saw in words in subsection 2.3.2 about collider bias in teh study of Sackett (Sackett, 1979). The DAG is depicted in 2.14.

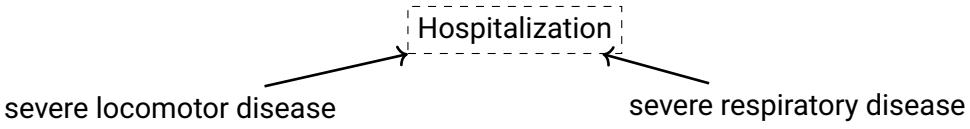



Figure 2.14: By only looking at hospitalized patients, which means adjusting for a collider, Sackett observed spurious correlation between the two disease groups. This is another example of the explain away effect. Once he looked at hospitalized and non hospitalized patients, was not adjusting for a collider anymore, he found independence, which is indeed the correct estimate from the true model.

### 2.6.2 Simpson's Paradox

Edward H. Simpson (1922-2019) described a phenomenon, currently known as Simpson's Paradox, in which one would find trends in subgroups of a population but the trend would reverse or disappear when the groups are combined (Simpson, 1951). A very famous example comes from a real-life study in which two treatments were investigated for kidney stones (Charig et al., 1986). What is surprising about the analysis initially performed is that treatment A (open surgical procedures) has a higher success rate in both patients with small (93% vs 87%) and large kidney stones (73% vs 69%), but treatment B (closed surgical procedures) has as higher success rate for everyone (83% vs 78%), which means all patients, not grouped by kidney stone size. How is this possible?

| | **Treatment A** | **Treatment B** |
|---|---|---|
| **Small stones** | Group 1 **93% (81/87)** | Group 2 87% (234/270) |
| **Large stones** | Group 3 **73% (192/263)** | Group 4 69% (55/80) |
| **Both** | 78% (273/350) | **83% (289/350)** |

In total, there were 700 patients in the study, and even though 50% of patients were assigned to each treatment arm, it's possible to see that the distribution of stone size, that correlates with disease severity, was unbalanced between treatments. About 77% of patients who received treatment B had small stones against only about 25% for treatment A, that is, the difficult cases were more prevalent for treatment A (75%) when compared to treatment B (23%). If you don't consider this, you will feel inclined to think treatment B is better, but that's because the light cases were assigned to this arm. Stone size is a confounder here and the appropriate way to estimate the effect of treatment on the outcome of patients is by adjusting for stone size as can be seen In Figure 2.15.

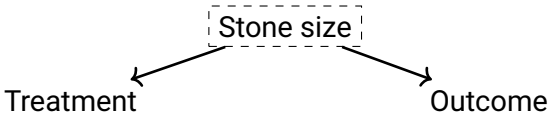
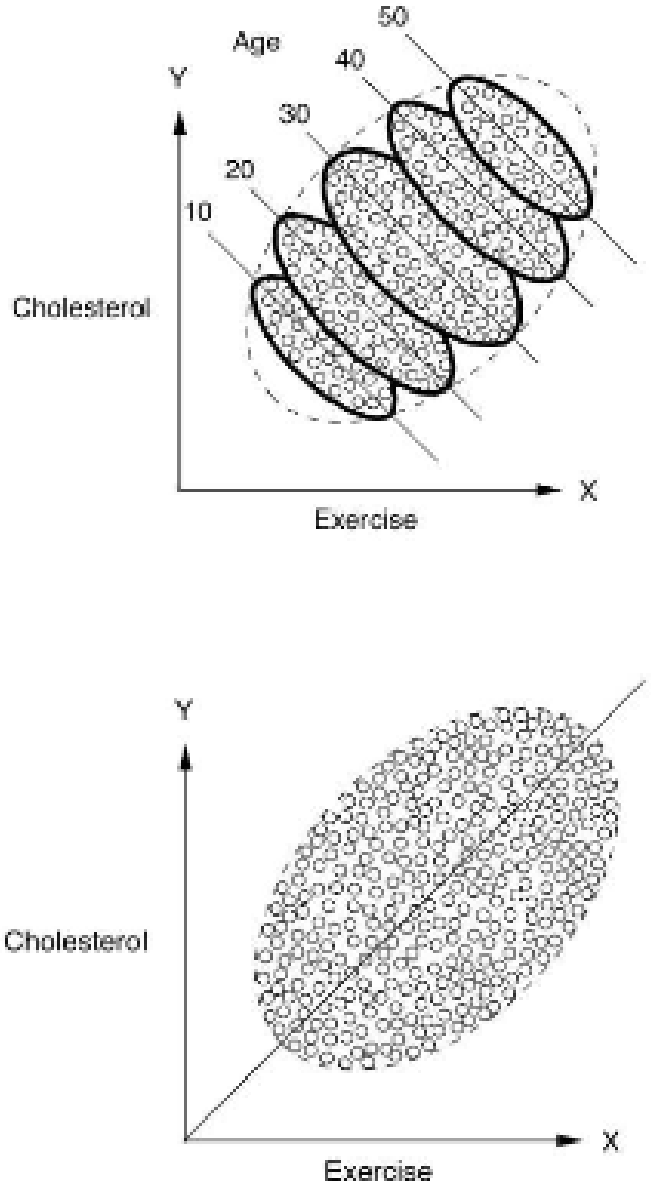



Figure 2.15: When adjusting for stone size, we see treatment A being better than treatment B.

Randomizing the participants of the study would remove the edge from stone size to treatment, which would give us a better distribution of stone sizes between treatment arms. One interesting visualization of Simpson's Paradox happening can be seeing in the Figure below from The Book of Why.

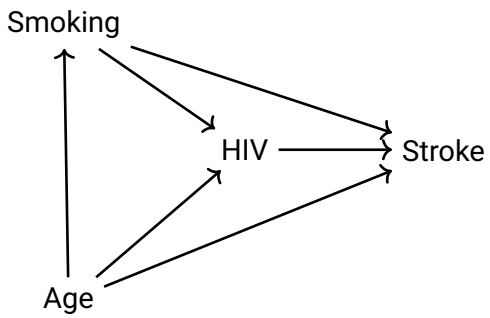



Retrieved from (Pearl and Mackenzie, 2018).

### 2.6.3 Table 2 Fallacy

Table 2 Fallacy occurs when multiple estimates are presented with different adjustments from a single model in a single table, leading the reader to misinterpret the causal relationship between the variables, as if they were all valid (Westreich and Greenland, 2013). Westreich and Greenland (2013) presents a real example retrieved from Madsen et al. (2011) in which several adjustments were made leading the reader to believe they're all valid hazard ratios, but the detail is that not all variables that were adjusted are confounders and therefore some relationships are just spurious, while others are valid causal estimates.

Another example given by Westreich and Greenland (2013) can be seen in the causal diagram below.




In the causal diagram above the relationship between the effect of human immunodeficiency virus (HIV) seroconversion and Stroke is confounded by Age and Smoking. One possible explanation for this is if the probability of infection with HIV increases with age and smoking, perhaps due to immunosuppression. A model adjusting for these confounders could be the logistic one in Equation 2.22.

$$logit(Stroke|HIV, Smoking, Age) = \beta_0 + \beta_1 \times HIV + \beta_2 \times Smoking + \beta_3 \times Age \tag{2.22}$$

One could report the estimated coefficients in a Table 2 leading many readers to assume $\beta_1$, $\beta_2$ and $\beta_3$ can be interpreted similarly and causally, after all the confounding was adjusted for! Nonetheless, even if the causal diagram is the true graph, the three coefficients represent different types of causal effects. $\beta_1$ is the conditional (at any given level of smoking and age) total effect of contracting HIV on Stroke. $\beta_2$, on the other hand, is the direct effect of Smoking relative to HIV. It is the fraction of the smoking effect on stroke that is not mediated through the smoking effect on HIV. $\beta_3$ has the same interpretation as $\beta_2$, what Westreich and Greenland (2013) called the controlled direct effect of aging (or smoking, for $\beta_2$) on stroke when the other

variables are held fixed, thus blocking the mediated effects through other variables.

Chapter 3

# iMIIC and iMIIC WebServer

> "Shallow men believe in luck or in circumstance. Strong men believe in cause and effect."
> Ralph Waldo Emerson (1803–1882)

The MIIC (Multivariate Information-based Inductive Causation) algorithm has been developed for many years now in Isambert lab at Institut Curie. It was based on the 3off2 scheme, a causal discovery algorithm for discrete datasets that provided a more robust approach to reconstruct graphical models from finite datasets, combining constraint-based and score-based approaches to infer structural independencies based on the ranking of their most likely contributing nodes (Affeldt and Isambert, 2016). MIIC has been extended numerous times to work with unobserved latent variables (Verny et al., 2017), to work with continuous and mixed-type data (Cabeli et al., 2020), to have guaranteed consistency regarding separation sets (Li et al., 2019), which contributed to interpretability, along with other features mentioned in more detail in this chapter, but also in chapters 5 and 6, which includes an improved algorithm for orienting edges.

Apart from the network reconstruction performed by the iMIIC algorithm, a web application has been developed, not only to make it easier for researchers to run iMIIC on their data off-premise (on the cloud), but also to provide extra pre and post-processing to the data (Sella et al., 2018). The web application outputs a rich graph along with many statistics and possibility of customization of the network.

This chapter will detail the iMIIC algorithm, which is the latest version also known as interpretable MIIC, and the latest version of the web application, iMIIC WebServer.

## 3.1 Network reconstruction

iMIIC is a hybrid approach for causal discovery, combining constraint-based and score-based methods for robust network structure learning. It starts with a complete graph and tests for independence, just like constraint-based methods, with the detail that it uses multivariate mutual information for that, *i.e.* mutual or 2-point (conditional) information, to be more precise. Besides, it makes use of scores, also based on Multivariate Information, to pick the best contributors for the separation set and also for orienting edges. Both tasks are based on 3-point (conditional) information terms. You can see the default iMIIC algorithm in Algorithm 1.

We've seen in subsubsection 2.5.3 that the PC algorithm iterates over all combinations of neighbors of the two nodes being tested for independence, until it runs out of neighbors or independence (be it pairwise or conditional) is achieved. iMIIC does this differently, by using the chain rule of conditional mutual information taking off the contribution of each putative contributor individually, according to the equation below.

$$I(X;Y|\{U_i\},Z) = I(X;Y) - I(X;Y;U_1) - I(X;Y;U_2|U_1) - ... - I(X;Y;Z|\{U_i\}) \tag{3.1}$$

There are mainly two improvements over PC that this modification provides. The first one is speeding up the first step to obtain the skeleton (final graph structure without the orientations), for it removes the combinatorial search. Besides, it also prevents spurious independence in a number of ways, by checking the contributors individually and removing the contributions according to the magnitude of each removal.

The score equation, or rank, $R(X,Y;Z|\{U_i\}) = min(P_{nv}(XYZ|\{U_i\}), P_b(XY|Z,\{u_i\})$, is the minimum value between the two conditions that indicate that Z really contributes to decreasing the Mutual Information found so far in $I(X;Y|\{U_i\}$. $P_{nv}(XYZ|\{U_i\})$ is the probability that $X - Z - Y$ is not a v-structure and is defined as:

$$P_{nv}(XYZ|\{U_i\}) = \frac{1}{1+e^{NI'(X;Y;Z|\{U_i\})}} \tag{3.2}$$

On the other hand, $P_b(XY|Z,\{U_i\})$ is the probability that $X - Y$ is the base and is defined as:

$$P_b(XY|Z,\{U_i\}) = \frac{1}{1+\frac{e^{-NI'(X;Z|\{U_i\})}}{e^{-NI'(X;Y|\{U_i\})}}+\frac{e^{-NI'(Y;Z|\{U_i\})}}{e^{-NI'(X;Y|\{U_i\})}}} \tag{3.3}$$

One new thing in the equations above is $I'$ (I prime), which is a regularized mu-

tual information using Normalized Maximum Likelihood (NML) to correct for finite sample size (Affeldt and Isambert, 2015), according to the equation below:

$$I'(X;Y|U) = I(X;Y|U) - \frac{k_{I(X;Y|U)}}{N} \tag{3.4}$$

$$I'(X;Y;Z|U) = I(X;Y;Z|U) - \frac{k_{I(X;Y;Z|U)}}{N} \tag{3.5}$$

The idea of regularization is to prevent iMIIC from picking complex models. View Chapter 6 for more details on Conservative MIIC, which is contained in iMIIC, and is an improvement of the previous use of regularized mutual information in MIIC.

**Algorithm 1** iMIIC network reconstruction algorithm

**Require:** $\mathcal{D}$
**Skeleton reconstruction**
$\mathcal{G} \leftarrow$ the complete graph
**for** all edges $X - Y \in \mathcal{G}$ **do**
**if** $I'(X;Y) < 0$ **then**
Delete edge $X - Y$ from $\mathcal{G}$
Sepset$\{X, Y\} \leftarrow \emptyset$
**else**
Find most contributing node $Z \in \{adj(X) \cup adj(Y)\}$ and compute $R(X, Y; Z|\{U_i\})$
**end if**
**end for**
**while** There is a link $X - Y$ with $R(X, Y; Z|\{U_i\} > \frac{1}{2})$ **do**
**for** Top link $X - Y$ with highest rank $R(X, Y; Z|\{U_i\})$ **do**
Expand contributing set $\{U_i\} \leftarrow \{U_i\} + Z$
**if** $I'(X;Y|\{U_i\}) < 0$ **then**
Delete edge $X - Y$ from $\mathcal{G}$
Sepset$\{X, Y\} \leftarrow \{U_i\}$
**else**
Find next most contributing node $Z \in \{adj(X) \cup adj(Y)\}$ and compute $R(X, Y; Z|\{U_i\})$
**end if**
Sort the rank list $R(X, Y; Z|\{U_i\})$
**end for**
**end while**
**Skeleton orientation**
Sort list of unshielded triplets $\mathcal{L}_c = \{(X, Z, Y)_{X \not\sim Y}\}$ in decreasing order of $|I'(X;Y;Z|\{U_i\})|$
**repeat**
Take $(X, Z, Y)_{X \not\sim Y} \in \mathcal{L}_c$ with highest $|I'(X;Y;Z|\{U_i\})|$ on which $R_0$ or $R_1$ orientation rules can be applied
**if** $I'(X;Y;Z|\{U_i\} < 0)$ **then**
if $(X, Z, Y)_{X \not\sim Y}$ has no diverging orientation, apply $R_0$ and orient $X \rightarrow Z \leftarrow Y$
**else**
if $(X, Z, Y)_{X \not\sim Y}$ has one converging orientation, apply $R_1$ and orient $X \rightarrow Z \rightarrow Y$
**end if**
Update all orientations of $(X, Z, Y)_{X \not\sim Y} \in \mathcal{L}_c$
**until** No additional orientation can be obtained
**return** $\mathcal{G}$

### 3.1.1 Consistency for Separation Sets

By the time the algorithm outputs the final graph, it's possible that the graph is different from an earlier version used to decide on edges and orientations. This has been initially observed in PC and fixed by Li et al. (2019). Imagine that at some point, an edge between $X$ and $Y$ was removed due to a common neighbor $Z$ ($X - Z - Y$), or a variable in a path between $X$ and $Y$. However, in the final graph, there is no path anymore, so that separation set is inconsistent to the skeleton observed in the final graph. Taking orientation into consideration, it could happen that a node that was in the path between $X$ and $Y$ was still in the path in the final graph, but this node is now a common descendant of $X$ and $Y$, a collider for example. This separation set is also inconsistent to the final oriented graph, as a common descendant must not be controlled to remove information between ancestors. This illustrated in Figure 3.1 and Figure 3.2.

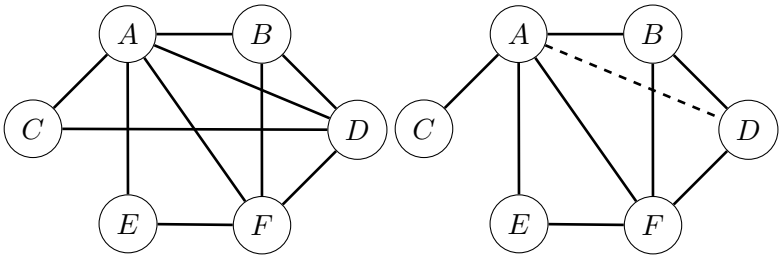



Figure 3.1: The graph on the left shows a point in a time in which there was a path between $A$ and $D$ through $C$ and $A \perp\!\!\!\perp D|\{C, F\}$. However, in the graph on the right, which is the final skeleton, it's possible to see that at some later moment the edge between $C$ and $D$ was removed, therefore $C$ is not in a path between $A$ and $D$ anymore, which makes the separation set $\{C, F\}$ inconsistent with the final graph.

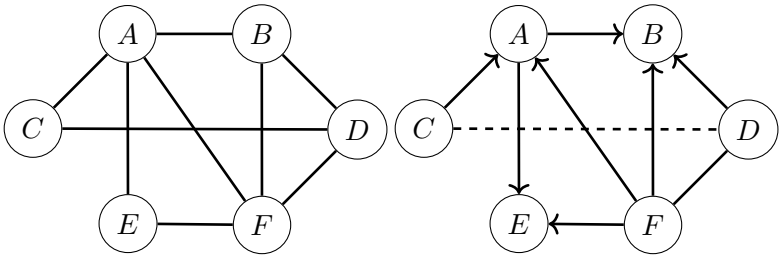



Figure 3.2: At some point, the edge between $C$ and $D$ was removed because $C \perp\!\!\!\perp D|B$. However, in the final graph it is seen that $B$ was not a mediator, but actually a collider, a common descent of $C$ and $D$ and therefore it should not be considered for the separation set. The separation set $\{B\}$ is thus inconsistent to the final oriented graph.

The separation set consistency check works in an iterative way. By the end of the iteration, the separation sets are checked as described in Figures 3.1 and **??**. On the second run, when deciding for the separation sets, the final graph of the previous iteration will be taken into consideration. Ideally, the goal should be to find two successive identical graphs, but there is no guarantee that this will necessarily happen, so what we expect is a sequence of graphs containing two identical graphs, which is given the name of a consistent cycle, that is, a series of graphs $G_{k-n}, G_{k-n+1}, ..., G_k$ such that $G_{k-n} = G_k$. The set $\{G_{k-n}, G_{k-n+1}, ..., G_k\}$ is then called a consistent cycle. The final consistent graph outputted by iMIIC is the union of the graphs in the aforementioned set, which is guaranteed to be consistent.

### 3.1.2 Latent confounder, putative and genuine orientations

In constraint-based algorithms for causal discovery, it's common to see orientation propagation. If we have a graph skeleton with two unshielded triples $X - Z - Y$ and $X - Z - T$, where $X - Z - Y$ is shown to be a v-structure ($X \rightarrow Z \leftarrow Y$) but not $X - Z - T$, then algorithms such as PC will orient the non v-structure as $X \rightarrow Z \rightarrow T$, because it's the only orientation compatible with a DAG. iMIIC brings the idea of orientation probabilities assigning a probability to every edge extremity which allows us to distinguish between a latent unmeasured confounder causing $X$ and $Y$ (visually depicted as $X \longleftrightarrow Y$), an edge with a probability suggesting an arrowhead in only one of the extremities, named putative edge ($\leftarrow$ or $\rightarrow$), and an edge for which there are probabilities suggesting an arrowhead in one of the extremities and an arrowtail in the other extremity, named genuine causal edge ($\Rightarrow$ or $\Leftarrow$). To decide on the orientation of every extremity, three basic rules could be considered:

1. If the probability $p$ is 0.5, the orientation for that extremity is undetermined.
2. If $p > 0.5$ the orientation for this extremity is likely an arrowhead.
3. If $p < 0.5$ the orientation for this extremity is likely an arrowtail.

However, iMIIC provides a parameter called orientation probability cut $p* \geq 0.5$ that filters probabilities in all extremities. Instead of the three rules above comparing always to $0.5$, in order for an extremity to be considered an arrowhead it must have probability $p > p*$ and to be an arrowtail $p < 1 - p*$, otherwise ($1 - p* < p < p*$) it's undetermined. With this in mind, the difference between a putative causal edge ($\rightarrow$ or $\leftarrow$) and a genuine causal edge ($\Rightarrow$ or $\Leftarrow$) is that in the putative case, $p > p*$ for one extremity (arrowhead) but $1 - p* < p < p*$ for the other extremity, our condition only

passes in one extremity, which has an arrowhead. The genuine causal edge case, having a certain belief that the tail is a tail and the head is a head gives us some reasonable confidence that this is a genuine causal edge.

## 3.2 iMIIC WebServer

iMIIC WebServer has evolved to be a powerful web application not only to run MIIC on uploaded datasets through a web browser, but also a powerful pipeline orquestration tool for the iMIIC R package. Most of the pre, post-processing and job management source code has been re-written in R and Python, together with some PHP, HTML, CSS and JavaScript code for the front and backend of the web application. The web application progressed along with the R package to show statistics and new visual features in the graph viewer, such as putative and genuine arrowheads, processing of information regarding individual contributions of the variables in the adjustment set, among other things. Figure 3.3 shows summary statistics for every edge (both removed, and retained). The user currently can see how much of the information between $X$ and $Y$ are left, and removed by which variables, or in the case of removed edges, which variables were responsible for removing the full information between the two variables. Figure 3.4, on the other hand, shows statistics about triplets, with information regarding the probabilities that led the final orientation observed in the graph viewer. Figure 3.5 shows a new tab for data dictionary, taking advantage of info that was already uploaded by the user for the job. In Figure 3.6 it's possible to see another new feature which is job comparison. The iMICC R package has numerous parameters that can be defined by the user. Sometimes, it's interesting to understand what changed between networks inferred with different parameters and to obtain some statistics between such networks. The feature allows users to view the union graph for the skeletons only or including orientation, intersection graph for the skeletons only or including orientation, among other set operations such as set difference.

The graph viewer comes together with many other tools that allow the user to customize the network, including node size, focus (change the transparency of other parts of the network), among other things.

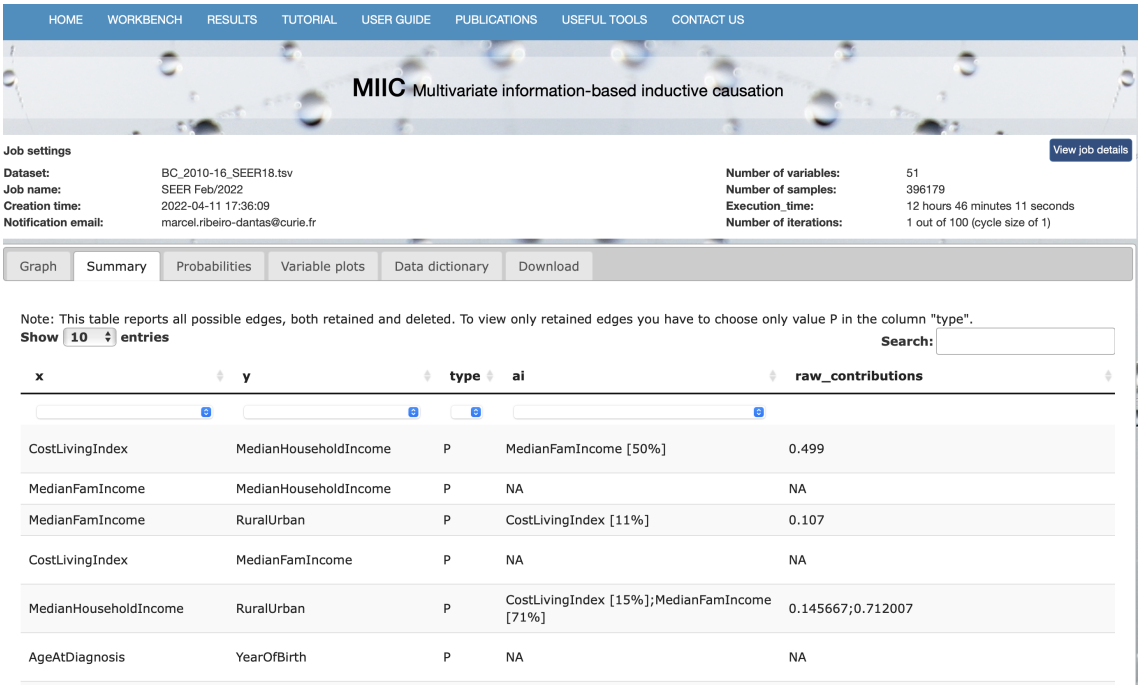
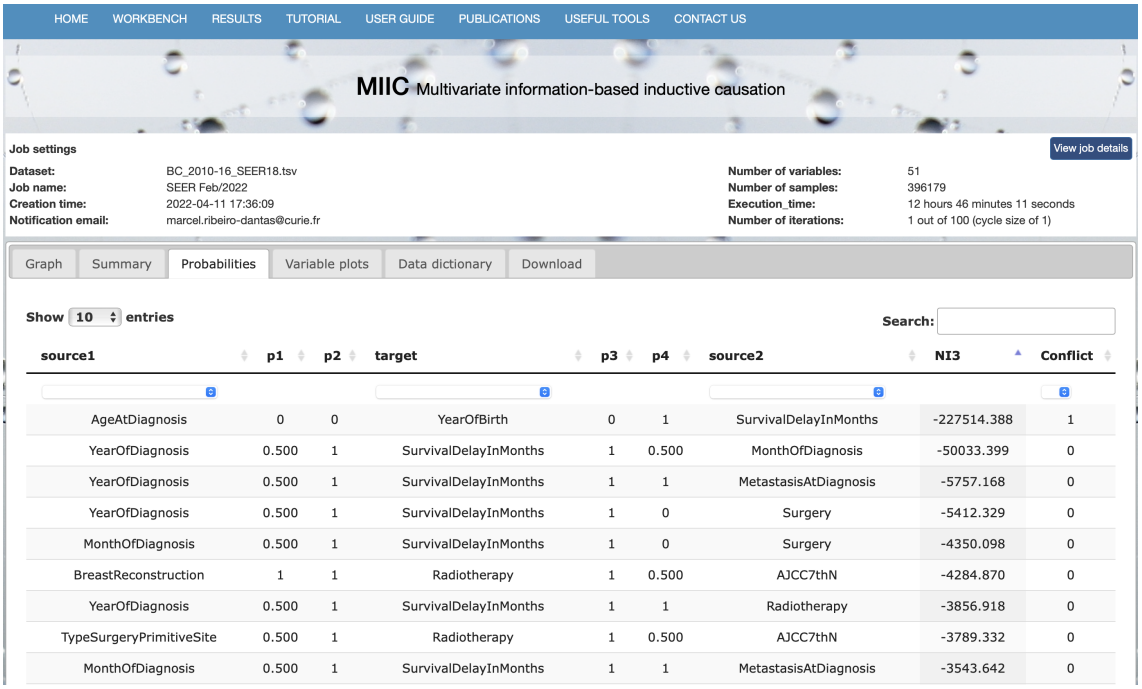
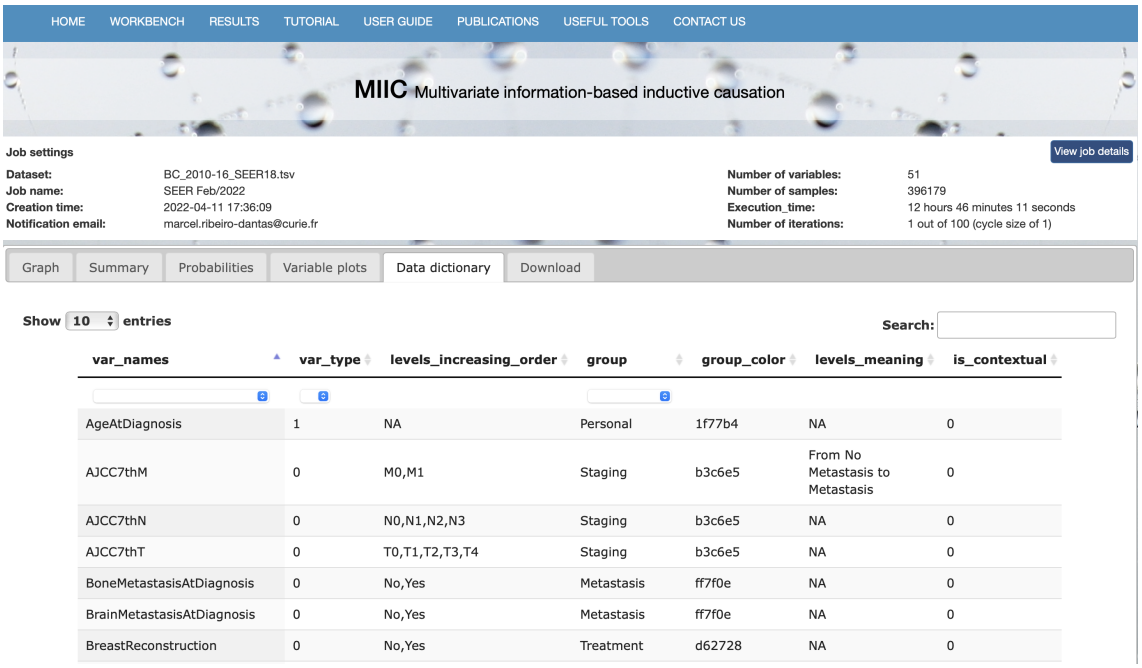



Figure 3.3: Summary statistics retrieved from https://miic.curie.fr/job_results.php?id=SEER2022. There are more statistics scrolling the window to the right.




| source1 | p1 | p2 | target | p3 | p4 | source2 | NI3 | Conflict |
|---|---|---|---|---|---|---|---|---|
| AgeAtDiagnosis | 0 | 0 | YearOfBirth | 0 | 1 | SurvivalDelayInMonths | -227514.388 | 1 |
| YearOfDiagnosis | 0.500 | 1 | SurvivalDelayInMonths | 1 | 0.500 | MonthOfDiagnosis | -50033.399 | 0 |
| YearOfDiagnosis | 0.500 | 1 | SurvivalDelayInMonths | 1 | 1 | MetastasisAtDiagnosis | -5757.168 | 0 |
| YearOfDiagnosis | 0.500 | 1 | SurvivalDelayInMonths | 1 | 0 | Surgery | -5412.329 | 0 |
| MonthOfDiagnosis | 0.500 | 1 | SurvivalDelayInMonths | 1 | 0 | Surgery | -4350.098 | 0 |
| BreastReconstruction | 1 | 1 | Radiotherapy | 1 | 0.500 | AJCC7thN | -4284.870 | 0 |
| YearOfDiagnosis | 0.500 | 1 | SurvivalDelayInMonths | 1 | 1 | Radiotherapy | -3856.918 | 0 |
| TypeSurgeryPrimitiveSite | 0.500 | 1 | Radiotherapy | 1 | 0.500 | AJCC7thN | -3789.332 | 0 |
| MonthOfDiagnosis | 0.500 | 1 | SurvivalDelayInMonths | 1 | 1 | MetastasisAtDiagnosis | -3543.642 | 0 |

Figure 3.4: Triplets statistics retrieved from https://miic.curie.fr/job_results.php?id=SEER2022. There are more statistics scrolling the window to the right.




| var_names | var_type | levels_increasing_order | group | group_color | levels_meaning | is_contextual |
|---|---|---|---|---|---|---|
| AgeAtDiagnosis | 1 | NA | Personal | 1f77b4 | NA | 0 |
| AJCC7thM | 0 | M0,M1 | Staging | b3c6e5 | From No Metastasis to Metastasis | 0 |
| AJCC7thN | 0 | N0,N1,N2,N3 | Staging | b3c6e5 | NA | 0 |
| AJCC7thT | 0 | T0,T1,T2,T3,T4 | Staging | b3c6e5 | NA | 0 |
| BoneMetastasisAtDiagnosis | 0 | No,Yes | Metastasis | ff7f0e | NA | 0 |
| BrainMetastasisAtDiagnosis | 0 | No,Yes | Metastasis | ff7f0e | NA | 0 |
| BreastReconstruction | 0 | No,Yes | Treatment | d62728 | NA | 0 |

Figure 3.5: Data dictionary retrieved from https://miic.curie.fr/job_results.php?id=SEER2022

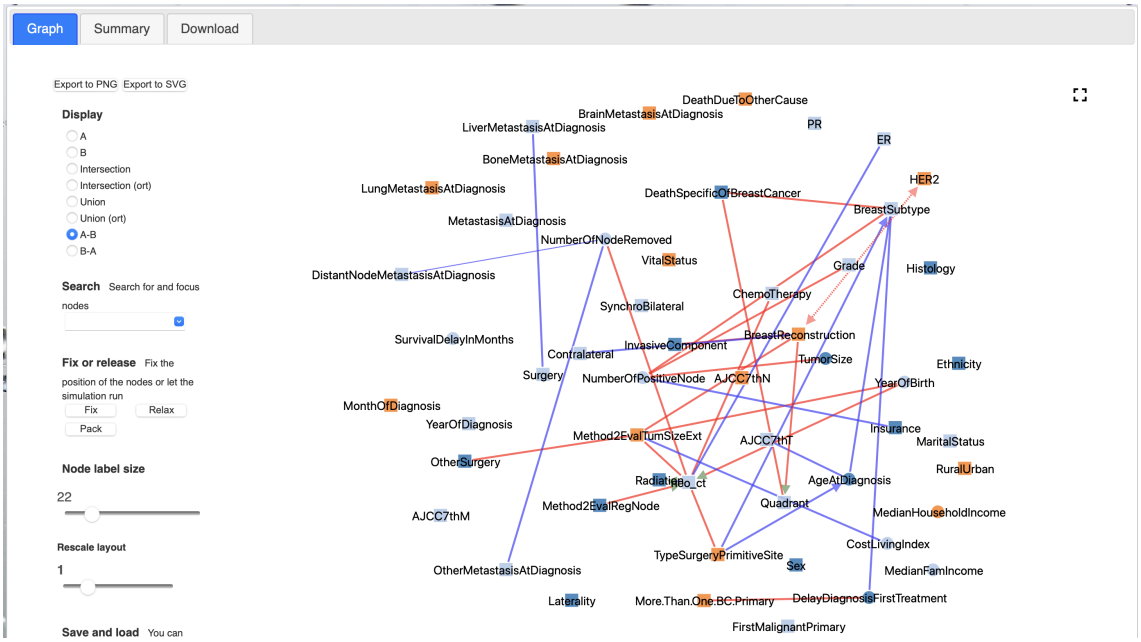



Figure 3.6: Job comparison retrieved from https://miic.curie.fr/

Chapter 4

# SEER

> "One of the first things taught in introductory statistics textbooks is that correlation is not causation. It is also one of the first things forgotten."
> Thomas Sowell (1930–)

This chapter gets into more detail about the Surveillance, Epidemiology, and End Results (SEER) Program.

## 4.1 SEER Program

The program is part of the National Cancer Institute of the United States and has been for a long time a source of epidemiological information when it comes to cancer. The program started collecting data on January 1st, 1973, in the states of Connecticut, Iowa, New Mexico, Utah, Hawaii, and the metropolitan areas of Detroit and San Francisco-Oakland. In more recent years it had expanded to major population centers in Georgia, Washington, Louisiana, New Jersey, Puerto Rico, Alaska, California, Kentucky (including Native American populations in Arizona, Alaskan Natives, and Hispanic populations in California) and in 2022 it has already reached the following 24 registries[1]: Alaska Native Tumor Registry, Arizona Indians, Cherokee Nation, Connecticut, Detroit, Atlanta, Greater Georgia, Rural Georgia, San Francisco-Oakland, San Jose-Monterey, Greater California, Hawaii, Idaho, Iowa, Illinois, Ken-

[1] Cancer registries receive and collect data about cancer patients in a region. There are two major types of cancer registries: population-based registries and hospital-based registries. Cancer registrars are the people who collect and report cancer data.

tucky, Los Angeles, Louisiana, Massachusetts, New Mexico, New Jersey, New York, Seattle-Puget Sound, Texas and Utah.

It currently covers approximately 48% of the U.S. population including 42% percent of Whites, 44.7% percent of African Americans, 66.3% percent of Hispanics, 59.9% percent of American Indians and Alaska Natives, 70.7% percent of Asians, and 70.3% percent of Hawaiian/Pacific Islanders. It is a widely used dataset with 17,000+ publications using SEER data for the primary analysis and 86,000+ referencnig SEER data as of 2022. So far, data on 11,000,000+ cases diagnosed with cancer have bene collected between 1973 and 2022.

## 4.2 SEER database

Data collected by some registries are not available for research, such as data on Arizona Indians and Cherokee Nation. Throughout the years, there were cases in which access to a variable would only be allowed for researchers with residence in the US. Variables that have been collected for a long time will have a lot of missing values, due to regions that started being covered only recently. Some variables stopped being collected, so more recent cases will have missing values. Many new variables also were only collected more recently, which means there will be a lot of missing values for patients diagnosed in the past. Thinking of all these limitations, and our interest in studying patients diagnosed with breast cancer, we decided to work on a subset of the SEER dataset for cases diagnosed between 2010 and 2016, since some interesting variables only started being collected after 2009 (2010+), such as Breast Subtype, metastasis site information, among other, and the latest version of the data released at the time had data until 2016.

Even though the SEER Program collects data on many different types of cancer, some extra care is given to breast cancer. Differently from most cancers, there are columns specific to Breast Cancer in the SEER database, such as Breast Subtype and data on Estrogen and Progesterone Receptor. This makes it a database even more interesting to use if researchers are focused on breast cancer, which is our case by having a reference hospital for breast cancer treatment and collaborators with expertise in the field.

## 4.3 Preprocessing

I was given access to the default "basic" SEER dataset when I had my account approved and, later, got my requests accepted for what the program calls Special-

ized Databases (Calculated Months Field from DX to Treatment Database, Census Tract-level SES and Rurality Database, and Treatment Database). In order to access all this, one needs to user the SEER*Stat software and through it I downloaded a dataset with X cases of patients diagnosed with cancer and Y mixed-type variables. Some variables do not contain values for the period 2010-2016, and some variables are specific to other types of cancer. Filtering out such variables brings the number of variables in the dataset to the number of Z.

There was a heavy step of preprocessing due to lack of standard in the values stored in the variables. Some variables have a value of "Blank(s)" for missing values, but other variables have alternate values for missing value such as: N/A, Unknown, NA, U, Unknown CHSDA, UNK Stage, Not applicable, Unknown or not applicable, among other. Other variables had codes that hindered the activity of transforming variables, doing analysis or variable engineering and thus before starting this step I checked the data dictionary and other sources of documentation to have actual information as values in the variables of interest.

After the steps aforementioned, there were 407,791 breast cancer records for the period 2010-2016, but only 396,179 distinct patients due to multiple breast primary tumors for some patients. For each patient, we selected the first breast primary tumor recorded in SEER and indicated the total number of breast cancer primaries during the 2010-2016 period in the variable MoreThanOneBCPrimary. SynchroBilateral was also engineered to identify patients who had tumors in both breasts diagnosed within less than 180 days of each other, while Contralateral identifies patients who had a subsequent tumor in the other breast diagnosed more than 180 days after the first breast tumor primary. Some categorical variables had some of their categories merged, either because these categories had the same general meaning or because they were too rare amongst patients (i.e. <0.1% of patients excluding those with missing data for the considered variable). These variables include Ethnicity, TypeSurgeryPrimitiveSite, Surgery, OtherSurgery, OtherMetastasisAtDiagnosis, Insurance and Histology. Hence, categories recorded in less than 0.1% of patients were merged and renamed to ‘Other’. BreastReconstruction was engineered based on TypeSurgeryPrimitiveSite (i.e. SEER surgery code ranges 43-49, 53-59, 63-69, and 73-75 were assigned ‘Yes’, while other surgery codes were assigned ‘No’). Radiotherapy was created from Radiation sequence with surgery, that has much fewer missing data (0.05%) than the original Radiation variable (49%). TumorSize merges two distinct variables that contained tumor sizes for years 2004-2015 and 2016+, respectively. Likewise, the largely missing 2016 information for the MetastasisAtDiagnosis variable was recovered based on information contained in spe-

cific metastasis variables (i.e. BoneMetastasisAtDiagnosis, LungMetastasisAtDiagnosis, LiverMetastasisAtDiagnosis, BrainMetas- tasisAtDiagnosis, OtherMetastasisAtDiagnosis). Finally, MedianFamIncome and MedianHouseHoldIncome are the average of these continuous variables over the periods 2007-2011, 2008-2012, 2009-2013, 2010-2014, 2011-2015, and 2012-2016.

Chapter 5

# Learning interpretable causal networks from very large datasets, application to 400,000 medical records of breast cancer patients

> "The fundamental activity of medical science is to determine the ultimate causation of disease."
> Wilfred Trotter (1872—1939)

This chapter includes the main contribution of this thesis which is a manuscript currently under consideration covering the progress and results obtained with causal discovery of interpretable causal networks applied to a subset of the SEER dataset.

# Learning interpretable causal networks from very large datasets, application to 400,000 medical records of breast cancer patients

Marcel da Câmara Ribeiro-Dantas[1,‡], Honghao Li[1,‡], Vincent Cabeli[1,‡], Louise Dupuis[1,‡], Franck Simon[1], Liza Hettal[1], Anne-Sophie Hamy[2,3,4], Hervé Isambert[1,*]

[1] CNRS UMR168, Institut Curie, Université PSL, Sorbonne Université, Paris, France
[2] INSERM U932, Institut Curie, Paris, France
[3] Department of Medical Oncology, Institut Curie, Saint-Cloud, France
[4] Department of Surgery, Institut Curie, Université Paris, Paris, France
[‡] these authors contributed equally to this work
[*] corresponding author information: herve.isambert@curie.fr, +33 1 56 24 64 74

**Discovering causal effects is at the core of scientific investigation but remains challenging when only observational data is available. In practice, causal networks are difficult to learn and interpret, and limited to relatively small datasets. We report a more reliable and scalable causal discovery method (iMIIC), that can learn interpretable causal networks for a wide range of biological, biomedical and presumably other complex heterogeneous data. It is based on a general mutual information supremum principle, which greatly improves the precision of inferred causal relations while distinguishing genuine causes from putative and latent causal effects. We showcase iMIIC on synthetic and real-life healthcare data from 396,179 breast cancer patients from the US Surveillance, Epidemiology, and End Results programme. More than 90% of predicted causal effects appear correct, while the remaining unexpected direct and indirect causal effects can be interpreted in terms of diagnostic procedures, therapeutic timing, patient preference or socio-economic disparity.**

Nationwide medical records contain massive amounts of real-life data on human health, including some personal, familial and socio-economic information, which frequently affect not only health conditions, but also timing of diagnosis, medical treatments and, ultimately, the survival of patients. Besides, such non-medical determinants of human health are usually controlled for in clinical trials, which select specific groups of patients through restrictive enrolment criteria. Yet, the wealth of information contained in real-life medical records remains largely under-exploited due to the lack of unsupervised methods and tools to analyze them without preconceived hypotheses. This highlights the need to develop new machine learning strategies to analyze healthcare data, in order to uncover unsuspected associations and possible cause-effect relations between all available information recorded in the medical history of patients, Fig. 1a.

Learning cause-effect relations from purely observational data has long been known to be, in principle, possible thanks to seminal works on causal discovery methods[1,2]. In essence, causal discovery infers cause-effect relations from specific correlation patterns involving at least three variables, which goes beyond the popular notion that pairwise correlation does not imply causation. However, while observational data account for the vast majority of available datasets across a wide range of domains, uncovering cause-effect relations still remains notoriously challenging in absence of systematic intervention, which might be impractical, too costly or unethical, when it concerns human health.

While causal discovery is tightly linked to methods designed to learn graphical models[1,2], most structure learning methods are not actually designed to uncover cause-effect relations. In particular, maximum likelihood approaches, such as Search-and-Score[3] or Graphical Lasso[4] methods, are restricted to specific model classes, assuming either fully directed graphs or fully undirected graphs, and cannot therefore learn the causal or non-causal nature of graph edges. By contrast, constraint-based causal discovery methods assume broader classes of graphs and can learn the orientation of certain edges solely based on observational data[1,2], Fig. 1b. To this end, they first learn structural constraints, in the form of conditional independence

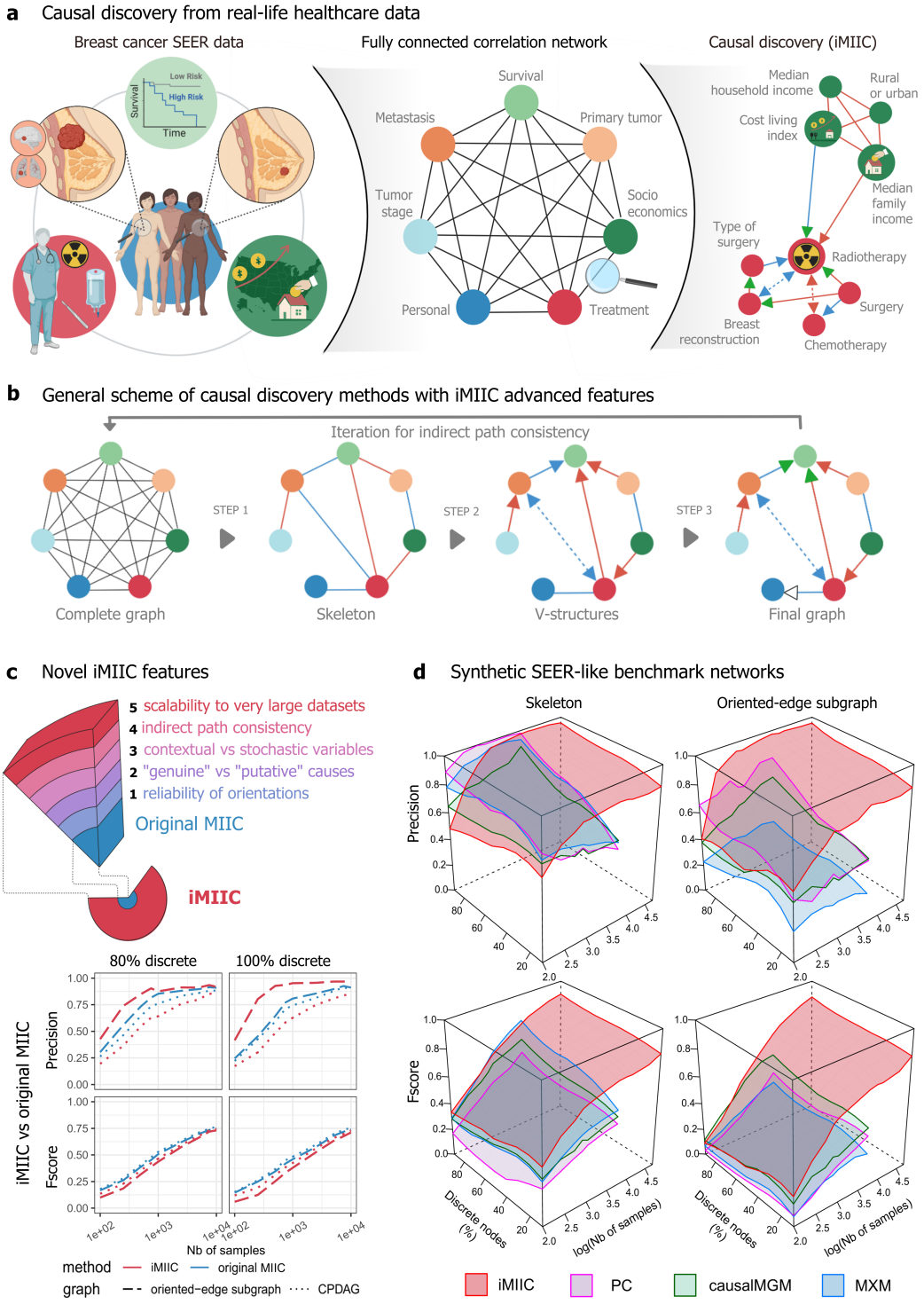



Figure 1: **Causal discovery from real-life healthcare data using constraint-based methods. (a)** SEER database includes 407,791 medical records of breast cancer patients diagnosed between 2010 and 2016. Causal discovery aims at uncovering cause-effect relations across such globally correlated datasets. **(b)** General scheme of constraint-based methods (including iMIIC's novel advanced features, see Methods): Step 1, removal of dispensable edges (guaranteeing indirect path consistency); Step 2, 'v-structure' orientation (with reliable orientations and latent common causes shown as bidirected edges); Step 3, propagation of orientation shown with white arrowhead (and distinction between 'putative' and 'genuine' causes, green arrowheads). **(c)** Novel iMIIC advanced features and benchmark comparison with original MIIC. **(d)** Synthetic SEER-like benchmark networks with different proportions of discrete variables, see text, Methods and Extended Data Figs. 4-6. Created with BioRender.com

relations, which provide indirect and somewhat cryptic information about possible causal relationships between observed as well as unobserved variables, as outlined in Box 1. Yet, despite being theoretically sound given unlimited amount of data[5], constraint-based methods remain unreliable and difficult to interpret on the relatively small datasets, they can handle in practice.

We report here the advanced causal discovery method, iMIIC (interpretable MIIC), that can learn more reliable and interpretable causal graphical models, as well as, handle much larger datasets (*e.g.* including a few hundred thousand samples). The novel iMIIC method expands and greatly improves the interpretability and scalability of the recent structure learning method, MIIC (Multivariate Information-based Inductive Causation), combining constraint-based and information-theoretic frameworks[6–8]. In short, iMIIC brings a number of advances, which greatly enhance its causal discovery performance on synthetic and real-life datasets of all scales. In particular, iMIIC (i) quantitatively improves the confidence in edge orientation, (ii) distinguishes "genuine" from "putative" causal relations (Box 1), (iii) distinguishes contextual from stochastic variables, (iv) enforces indirect path consistency and quantifies their information contributions, and, finally, (v) enables scalability to very large datasets. These augmented capacities, which rely on conceptual advances and extensive algorithmic refactoring, are applied to reconstruct an interpretable causal network from the analysis of more than 400,000 medical records of breast cancer patients from the Surveillance, Epidemiology, and End Results (SEER) programme[9], Extended Data Fig. 1.

## Overview and limitations of causal discovery methods

Constraint-based causal discovery methods proceed through successive steps, outlined in Fig. 1b. The first step consists in removing, iteratively, all dispensable edges from an initial fully connected network, whenever two variables are independent or conditionally independent given a so-called separating set of conditioning variables. The second step then consists in orienting some of the edges of the undirected graph (named skeleton) to form so-called "v-structures", $X \rightarrow Z \leftarrow Y$, which are the signature of causality in observational data, Box 1. Finally, the third step aims at propagating the orientations of v-structures to downstream edges, Fig. 1b. However, traditional constraint-based methods lack robustness on finite datasets, as their long series of uncertain decisions lead to an accumulation of errors, which limit the reliability of the final networks. In particular, spurious conditional independences, stemming from coincidental combinations of conditioning variables, lead to many false negative edges and, ultimately, limit the accuracy of inferred orientations. The recent machine learning method, MIIC[6,8], learns more robust causal graphical models by first collecting iteratively significant information contributors before assessing conditional independences (see Methods). In practice, MIIC's strategy limits spurious conditional independences and greatly improves the sensitivity or recall (*i.e.*, the fraction of correctly recovered edges) compared to traditional constraint-based methods, Extended Data Figs. 2 and 3. In addition, MIIC can handle heterogeneous data (*i.e.* combining continuous and categorical variables) and missing data[8], as well as, unobserved latent variables[6], that are ubiquitous in many real-life applications.

Yet, the original MIIC method still presents a number of limitations, such as a lower reliability in predicting edge orientation than edge presence, that the novel iMIIC method aims to overcome, as outlined below. In practice, iMIIC is shown to greatly enhance the reliability, interpretability and scalability of causal discovery from large scale synthetic data, as well as, real-life observational datasets.

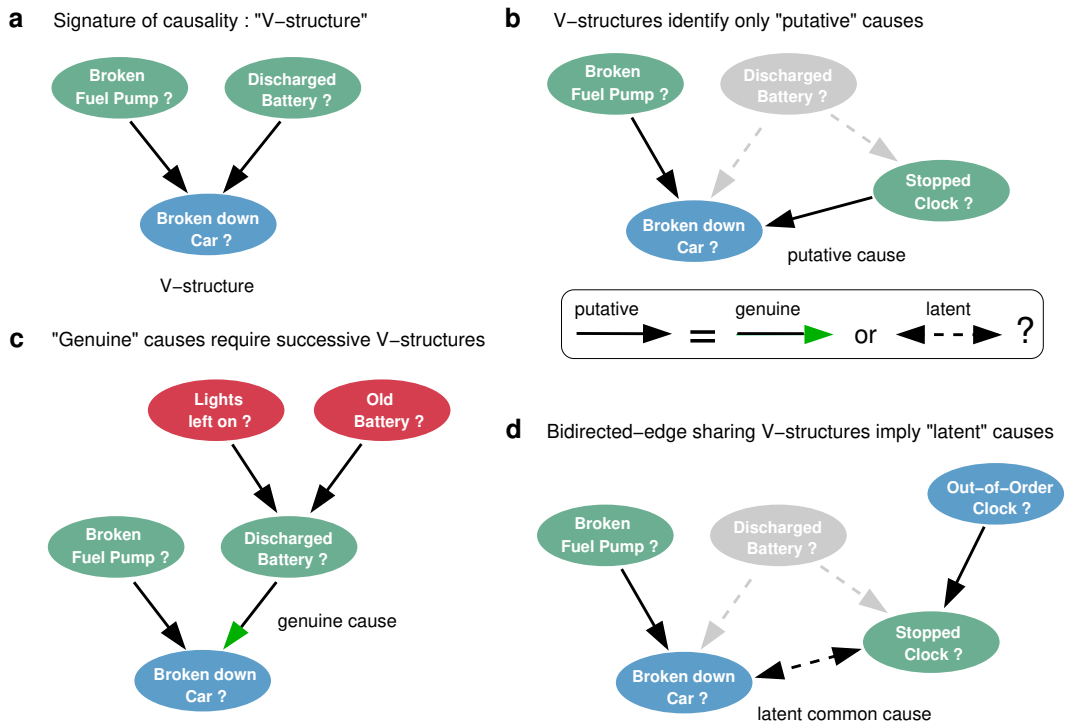
## Novel features of the advanced iMIIC method

**iMIIC improves the reliability of inferred orientations.** While the original MIIC significantly outperforms traditional constraint-based methods in inferring reliable orientations, a substantial loss in precision usually remains between MIIC skeleton and oriented graph predictions, Extended Data Fig. 3. This is due to orientation errors originating mainly from inconsistent v-structures, $X \rightarrow Z \leftarrow Y$, whose middle node $Z$ could also be included in the separating set of the unconnected pair $\{X, Y\}$, in contradiction with the head-to-head meeting of the v-structure. Inconsistent v-structures are particularly common for datasets including discrete variables with (too) many levels. To prevent such inconsistent orientations, iMIIC implements more conservative orientation rules, based on a general mutual information supremum principle[14,15] regularized for finite datasets, see Methods. In practice, it greatly enhances the reliability of predicted orientations with only a small sensitivity loss compared to MIIC original orientation rules, Fig. 1c.

## Box 1. Causal discovery principles from observational data: putative, genuine and latent causes.

We outline here the principles to uncover cause-effect relations in a purely observational dataset and distinguish "genuine" causes from "putative" and "latent" causes. The rationale is illustrated on the causally intuitive toy example of an imaginary dataset of old cars. (**a**) The signature of causality in such observational datasets corresponds to 3-variable "v-structure" subgraphs involving two *independent* and thus *unconnected* possible causes, "Broken fuel pump?" and "Discharged battery?", and a resulting effect, "Broken down car?". The converging orientations of this v-structure towards its middle variable, "Broken down car?", stem from the fact that these two edges cannot be undirected, nor can they point towards either "Broken fuel pump?" or "Discharged battery?", as these alternative graphical models would imply correlations contradicting the independence between "Broken fuel pump?" and "Discharged battery?". Alternatively, causal relations can sometimes be uncovered between two variables only, under the specific assumption of continuous additive noise models[10]. However, in the general case, causal discovery requires at least three and often more variables, as the independence between possible causes in a v-structure is frequently conditional on other variable(s), not considered here, defining a separating set, see Methods. Conversely, conditioning on the tip of a v-structure, here "Broken down car?", induces spurious associations between its independent possible causes[1,2]. Likewise, selecting a dataset with specific values for this tip variable results in spurious associations due to selection bias in the dataset[11–13], such as some apparent anti-correlation between different possible causes, "Broken fuel pump?" and "Discharged battery?", if only "Broken down car? = yes" are selected. (**b**) However, v-structures remain in fact causally ambiguous[2] as they only identify "putative" causes, which can either be "genuine" causes, displayed with a green arrowhead, or suggest the presence of unmeasured confounders, *i.e.* latent common causes unobserved in the dataset and represented with a bidirected edge. For instance, the variable "Clock stopped?", frequently used as a proxy for "Discharged battery?", also forms a similar v-structure with "Broken fuel pump?"; yet, it is well known that "Clock stopped?" cannot be a genuine cause of "Broken down car?", as tampering with a car's clock cannot actually cause a car to break down. (**c**) In absence of background knowledge and direct intervention on variables, showing that "Discharged battery?" is indeed a genuine cause of "Broken down car?" requires to exclude the possibility of an unobserved common cause (*i.e.* an unmeasured confounder) between "Discharged battery" and "Broken down car?". To this end, one needs to find another v-structure upstream of "Discharged battery?" (*e.g.* "Lights left on?"→"Discharged battery?"←"Old battery?") or to have prior knowledge about an upstream (putative) cause and to show that the effect of at least one upstream variable on the downstream variable "Broken down car?" is entirely *indirect* and mediated (at least in part) by the intermediary variable "Discharged battery?". This requires to find a conditional independence between an upstream variable and "Broken down car?" conditioned on a separating set, which includes the intermediary variable "Discharged battery". (**d**) Conversely, ruling out a putative cause as genuine cause requires to show that the relation actually originates from an unobserved common cause by finding a fourth variable (*e.g.* "Out-of-order clock?") defining another v-structure, inducing a bidirected edge between "Broken down car?" and "Clock stopped?" with the v-structure in (b).

The advanced iMIIC method distinguishes genuine from putative causal edges, as well as, undirected and bidirected edges, by assessing separate head or tail orientation probabilities at each edge extremity (see Results and Methods).



**iMIIC distinguishes "genuine" from "putative" causal relations.** Traditional constraint-based methods and indeed the original MIIC method merely discover "putative" causal relations, as v-structure orientations are actually compatible with both genuine cause-effect relations and the effects of unobserved common causes, as outlined on an intuitive example in Box 1. By contrast, iMIIC distinguishes "genuine" from "putative" causal edges by ruling out the effect of an unobserved common cause (or unmeasured confounder) for each predicted genuine causal edge. It is achieved by assessing separate probabilities of arrow head and tail for all oriented edges, see Methods. Genuine causal edges (represented with a green arrow head) are then predicted if both arrow head and tail probabilities are statistically significant, while causal edges remain "putative" if their tail probability is not statistically significant or cannot be determined from purely observational data. Likewise, bidirected edges, interpreted as the effect of unobserved common causes, correspond to two significant head probabilities, while all other cases are graphically represented as undirected edges.

**iMIIC distinguishes contextual from stochastic variables.** The separate probabilistic framework of arrow head *versus* tail orientations implemented in iMIIC also allows to include prior knowledge about certain head or tail orientations. For instance, including a few contextual variables in graphical models can help specify a control parameter or experimental conditions or characterize the personal profile of patients (*e.g.* sex, year of birth), depending on the nature of the dataset. Unlike most other variables of the dataset, such contextual variables are not stochastically varying and should have, by assumption, all their edges without incoming arrow head, *i.e.*, $p_{\text{tail}} = 1$. This expresses our prior knowledge that contextual variables cannot be the consequence of other observed or non-observed variables in the dataset.

**iMIIC enforces indirect path consistency and quantifies their information contributions.** The rationale behind the removal of dispensable edges in the first step of constraint-based causal discovery methods is that all statistical associations between disconnected variables should be graphically interpretable in terms of indirect paths in the final network. However, this is frequently not the case in practice[16]. In particular, there is no guarantee that the separating sets identified during this iterative removal of edges remain consistent in terms of indirect paths in the final network. To this end, iMIIC adapts a novel algorithmic scheme[16] to ensure that all separating sets identified to remove dispensable edges are consistent with the final inferred graph. It is achieved by repeating the constraint-based structure learning scheme, iteratively, while selecting only separating sets that are consistent with the skeleton or the partially oriented graph obtained at the previous iteration, as outlined in Fig. 1b. This indirect path consistency improves the interpretability of iMIIC inferred networks in terms of indirect effects, which are also quantified through indirect information contributions, see Methods.

**iMIIC outperforms existing methods on synthetic benchmark datasets.** The performance of iMIIC has been benchmarked against original MIIC as well as other state-of-the-art constraint-based methods on benchmark datasets with different proportions of discrete variables, see Methods. Fig. 1c demonstrates that iMIIC significantly improves the precision of orientations to the expense of a relatively small loss in orientation sensitivity and F-score for SEER-like benchmark datasets with large proportions of discrete variables. For instance, for $N = 500$, orientation precision (resp. F-score) already exceed 85% (resp. 32%) with iMIIC *versus* 73% (resp. 39%) with original MIIC, for SEER-like benchmark datasets with 80% discrete variables, and even 93% (resp. 25%) *versus* 64% (resp. 35%) for fully discrete datasets, Fig. 1c. In addition, iMIIC greatly outperforms the reliability and sensitivity of inferred orientations against other state-of-the-art constraint-based methods, Fig. 1d and Extended Data Figs. 4-6. In particular, iMIIC's orientation F-scores are about twice as high as PC algorithm[17,18]'s orientation F-scores, for all sample sizes and discrete variable proportions in these SEER-like datasets. For instance, for benchmarks with 80% discrete variables as in the actual SEER dataset, iMIIC already reaches 88% (resp. 44%) in precision (resp. F-score) for $N = 10^3$, against about 60% (18%) for conservative PC[18,19], 50% (36%) for causalMGM[20] and 24% (18%) for MXM[21]. For $N = 10^4$, iMIIC reaches 92% (73%) in precision (F-score), against about 75% (40%) for conservative PC, 62% (55%) for causalMGM and 30% (30%) for MXM. Finally, iMIIC reaches more than 90% for both orientation precision and F-score, for $N = 10^5$, which is beyond the sample size attainable by other methods. See Methods for comparisons with higher proportion of continuous variables.

## Application to nationwide medical record data

**SEER breast cancer data.** We applied iMIIC on a large breast cancer dataset[9] from the Surveillance, Epidemiology, and End Results (SEER) programme of the National Cancer Institute, which collects data on cancer diagnoses, treatment and survival for ∼35% of the US population, Fig. 1a. Breast cancer[22] is the most common invasive cancer in women and is curable in only 70-80% of patients with large disparities in terms of tumor subtypes and stages at diagnostic, initial and subsequent treatments, as well as patient's age, ethnicity, genetic predisposition, lifestyle or socio-economic situation. Numerous retrospective association studies[23–25] and a few causal inference investigations[26–29] have been reported on SEER's cancer data, making it a unique benchmark resource to assess the actual performance of causal discovery methods on real-life healthcare data.

**Robust iMIIC causal discovery analysis on ∼400,000 breast cancer patients.** We present here iMIIC's causal discovery analysis on SEER breast cancer data for the period 2010-2016. There are 407,791 medical records but only 396,179 distinct patients due to multiple breast primary tumors for some patients. Fifty one clinical, socio-economic and outcome variables have been selected for their relevance to breast cancer and for their limited redundancy or missing information, Extended Data Fig. 1. The resulting breast cancer network, Fig. 2a, provides an interpretable graphical model including 280 edges, for which most cause-effect relations are either known or can be ruled out based on common or expert knowledge as well as clinical practice. This assessment indicates that about 90% of genuine or putative causal effects inferred by iMIIC are correct, based on clinical and epidemiological knowledge, while an additional 8% of cause-effect relations seem plausible (see Supplementary Table 1). Besides, none of the predicted genuine causal edges connect pairs of non-cancer-specific variables, such as personal or socio-economic information, that are susceptible to a possible selection bias[11–13] through breast cancer diagnosis (Box 1). In addition, unmeasured (latent) confounders can be ruled out for genuine causal edges (Box 1) while contributions by measured confounders are estimated as indirect path contributions (see Methods). Yet, other sources of bias in data collection and analysis have been reported on the SEER database[30,31] (as discussed in the following section). This ∼400,000 patient clinical network is also robust to sub-sampling as it includes 90% of the edges of three networks learned from three independent subsets of 100,000 patients, Fig. 2b. In addition, 88% of the edge orientation probabilities are compatible between the three 100,000 patient subset networks and 92% of those are also compatible with the edge orientation probabilities of the full network (see Supplementary Table 1).

## Causal interpretation of iMIIC breast cancer network

We now address the clinical and socio-economic interpretation of the SEER breast cancer network inferred by iMIIC, Fig. 2a. We will focus, in particular, on the expected as well as more surprising genuine causal relations uncovered by iMIIC, and will propose interpretations of the counter-intuitive cause-effect predictions in terms of care pathway, therapeutic decisions, patient preferences or socio-economic determinants of healthcare. We present these results from the perspective of different classes of variables and associated subnetworks, starting with the survival subnetwork, then the primary tumor subnetwork, the surgery and subsequent treatment subnetwork, and finally the socio-economic subnetwork.

**Survival subnetwork.** The full network, Fig. 2a, contains four nodes associated with patient survival status at the end of 2016 and defining a survival subnetwork, that includes all variables directly linked to patient survival status, Fig. 3a. Beyond the vital status of each patient (dead or alive), two additional nodes specify the cause of death, either from breast cancer or from any other cause, and a third continuous variable corresponds to the survival or follow-up delay in months, subjected to the censoring period 2010-2016 of the study. Fig. 3a shows that known factors responsible for the death due to breast cancer are correctly recovered by iMIIC, such as metastatis at diagnosis (overall mortality rate 49.2%), with the worse distant metastases at diagnosis (brain and liver) also retaining direct links to both Death specific to breast cancer and Vital status, which accounts for their excess mortality rates, *i.e.* brain metastasis (70.5%) and liver metastasis (59.5%). Similarly, the number of metastasis-positive lymph nodes and the staging variables (AJCC7th T, N, and M) are all correctly connected to both death specific to breast cancer and vital status, and not to any other cause of death. By contrast, iMIIC infers causal relations between year of birth and death due to other cause, as well as, year of birth and vital status, as expected. We can also note that the deaths of patients, irrespective of their cause, are rightly predicted to lead to a reduction in their survival delays. Yet, Fig. 3a

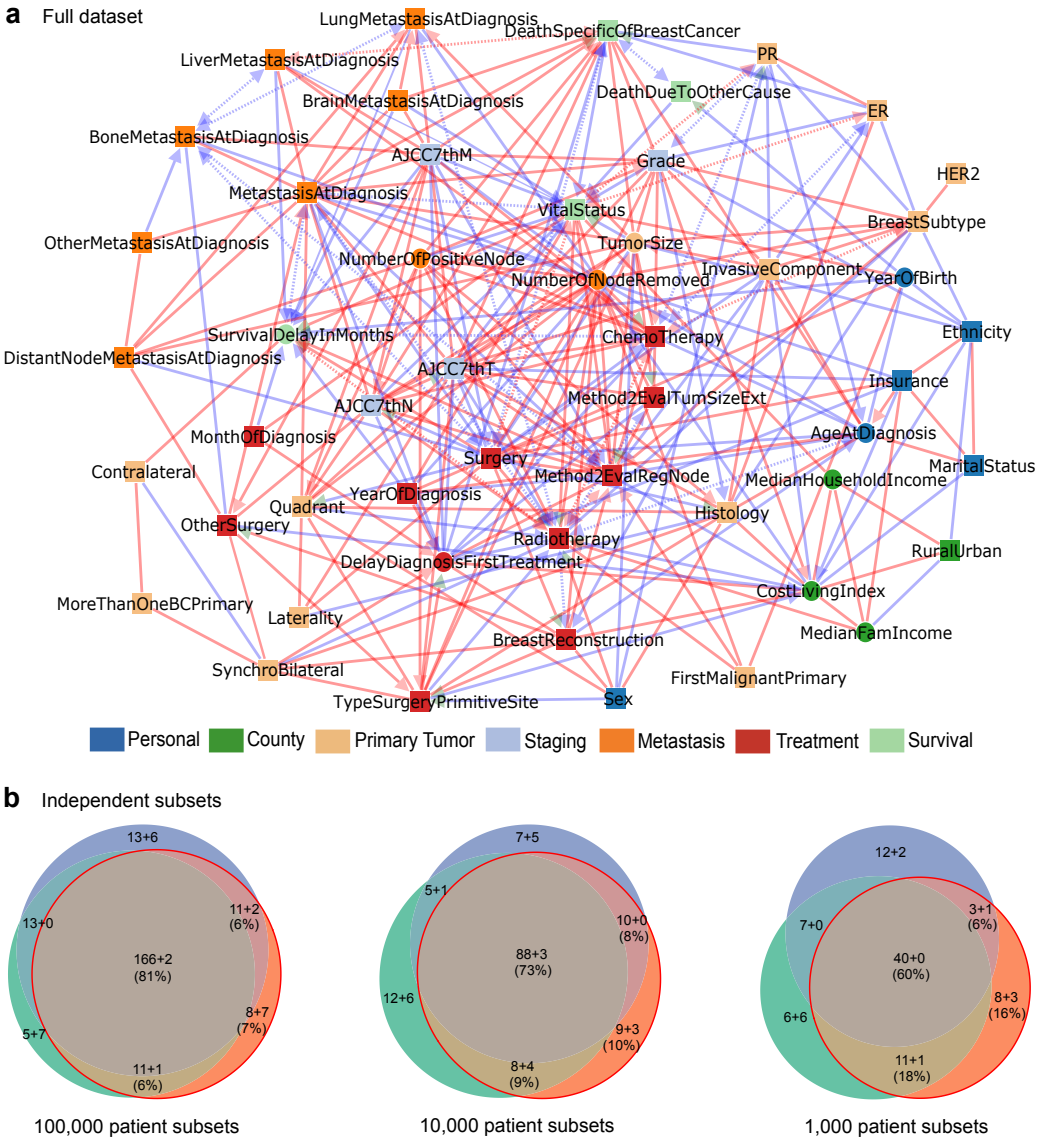



Figure 2: **SEER breast cancer networks inferred by iMIIC**. **(a)** The 51 node network inferred by iMIIC from SEER dataset including 396,179 breast cancer patients diagnosed between 2010 and 2016. This skeleton consistent network contains 280 edges and includes 2 contextual variables, Sex and Year of birth. The corresponding orientation consistent network contains 340 edges, Extended Data Fig. 7. See Supplementary Table 1 for a list and causal nature of each edges predicted by iMIIC. **(b)** Comparisons of networks inferred from three independent sub-samplings of the same size of 100,000, 10,000 or 1,000 patient subsets (from left to right). Number of shared edges (regardless of orientations) in the Euler diagrams are given as a sum $a + b$ where $a$ (resp. $b$) corresponds to the number of edges included in (resp. absent from) the full dataset network in (a). Percentages refer to the subset network with the median total number of edges (red circle).

contains also less intuitive findings. In particular, vital status is robustly inferred to 'cause' radiotherapy, both in the full dataset and in all three 100,000 patient subsets, with 51% of alive patients having undergone radiotherapy against only 27% of dead patients, Fig. 3b. This suggests that early death within the first few months after diagnosis may prevent radiotherapy for some patients who might have otherwise received this treatment, have they lived longer. This short term causal effect between vital status and radiotherapy is consistent with the rapid decline of the survival delay distribution for the first 3-6 months in absence of radiotherapy, Fig. 3c, which corresponds to the typical range of delays for radiotherapy after diagnosis, depending on whether it is performed as second treatment after surgery or as third treatment after both

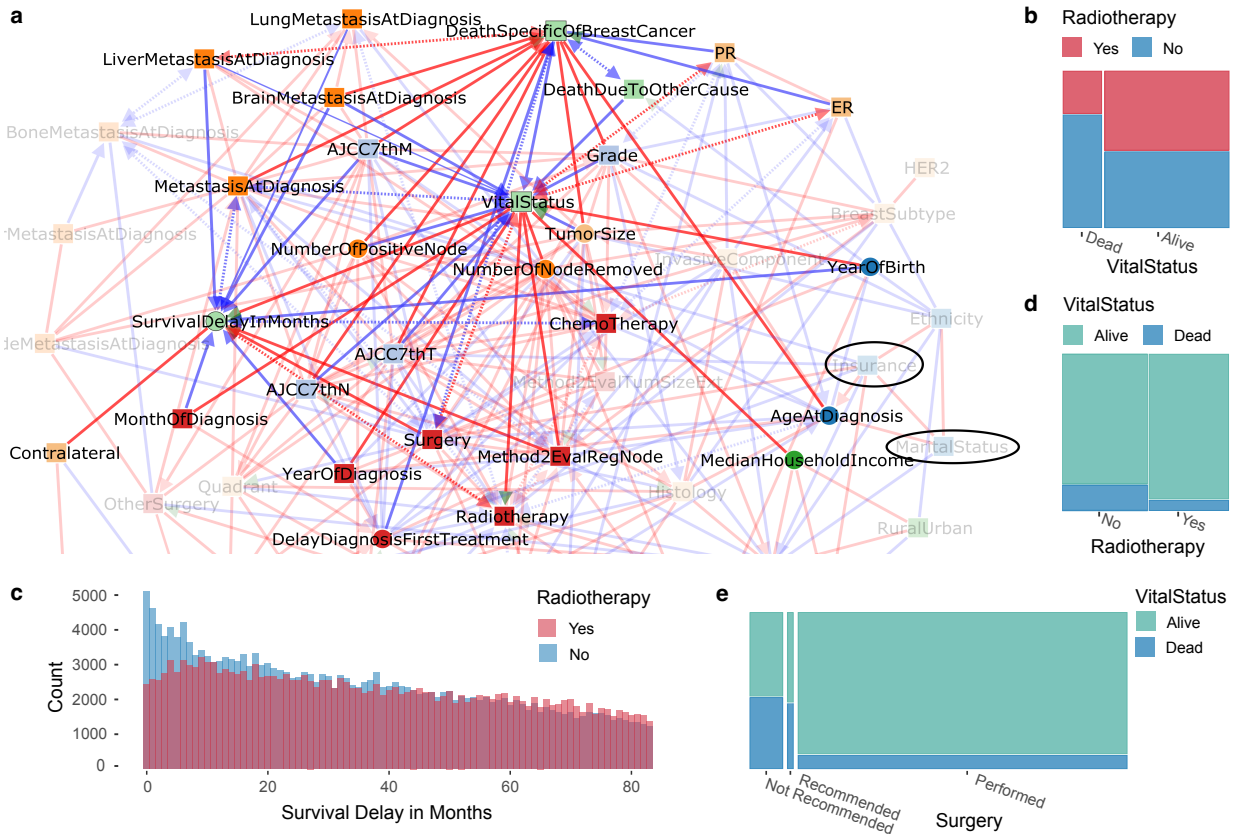



Figure 3: **Survival subnetwork inferred by iMIIC from SEER breast cancer dataset**. **(a)** Subnetwork highlighting direct relations with survival variables (VitalStatus, DeathSpecificOfBreastCancer, DeathDueToOtherCause, SurvivalDelayInMonths). The absence of direct links with other variables (such as Insurance and Marital Status highlighted in the network) can be interpreted in terms of indirect path contributions consistent with the network skeleton, see main text and Methods. **(b)** Joint distribution of Radiotherapy and Vital Status highlighting the counter-intuitive causal relation between them, see text. **(c)** Histogram of Survival Delay In Months for patients having received Radiotherapy or not. Each bin represents one month. The early blue peak suggests that a number of patients died within 3 to 6 months after diagnosis, hence, before they could receive Radiotherapy, in agreement with the causal direction predicted in **(a)**. This results in an over-estimated apparent benefit of Radiotherapy in **(d)**, see main text. **(d)** Joint distribution of Vital Status and Radiotherapy. **(e)** Joint distribution of Vital Status and Surgery.

surgery and chemotherapy[32]. All in all, this short term causal effect of vital status on radiotherapy outweighs the causally reversed, beneficial effect of radiotherapy on the long term survival of patients. This suggests a strong "immortal time bias"[30] in the apparent benefit of radiotherapy, Fig. 3d, which would need to be corrected with the "landmark method"[30,33] excluding patients dying within a specified period after surgery, or by emulating a target trial from observational data[34]. By contrast, surgery –which is typically performed within 5 to 8 weeks after diagnosis– is found to be the primary cause leading to the prolonged survival delay of patients, as discussed below, Fig. 3e and Fig. 4a.

Finally, we note that a number of variables that have been reported to be associated to survival variables are in fact indirectly rather than directly connected to them. This is, in particular, the case of insurance[35,36] and marital status[37,38]. The indirect effect of Insurance (with uninsured / medicaid / non-medicaid as categories) on Death due to breast cancer is shown to be indirectly explained through Surgery (50%), ChemoTherapy (14%), MaritalStatus (20%), Radiotherapy (9%), and Breast reconstruction (7%), see Eq. 10 in Methods. Similarly, the indirect effect of marital status (with single / married / separated / divorced / widowed categories) on Death due to breast cancer is shown to be indirectly explained through Surgery (58%), Year of birth (40%), and Ethnicity (2%).

**Primary tumor subnetwork.** Besides metastasis at diagnosis, the hormone receptor (ER/PR) status and the size of the primary tumor are also found to directly affect the vital prognosis of patients, Extended

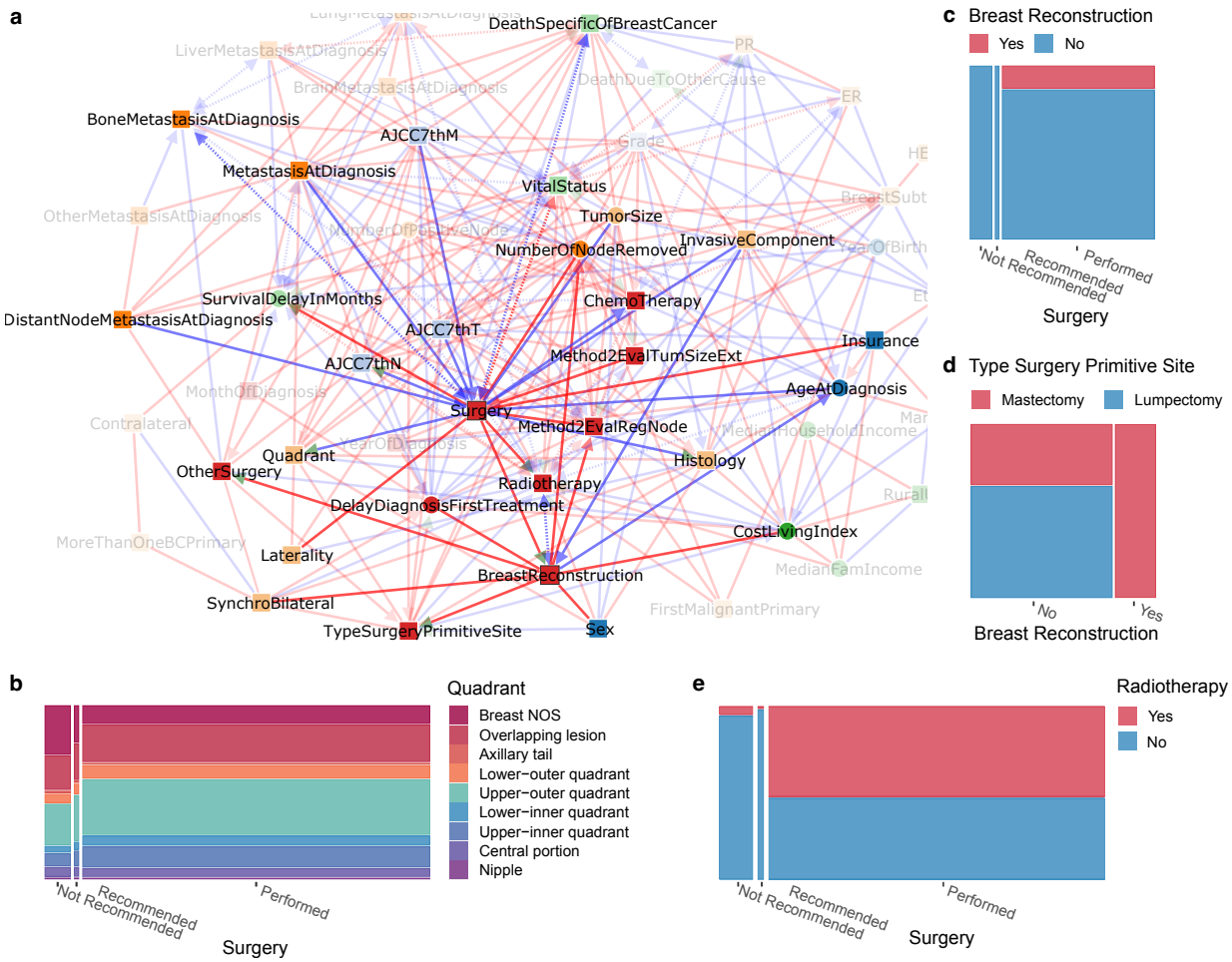



Figure 4: **Surgery and subsequent treatments subnetwork inferred by iMIIC from SEER breast cancer dataset**. **(a)** Subnetwork highlighting direct relations with Surgery and Breast Reconstruction. **(b)** Joint distribution of Quadrant and Surgery. **(c)** Joint distribution of Breast Reconstruction and Surgery. **(d)** Joint distribution of Type Surgery Primitive Site and Breast Reconstruction. **(e)** Joint distribution of Radiotherapy and Surgery. See main text for causal interpretation of the role of Surgery on refining primary tumor characterisation and subsequent therapeutic decisions including personal choice of patients.

Data Fig. 8a. In particular, iMIIC infers that ER status reduces the risk of death due to breast cancer from 17.7% (ER-) to 5.4% (ER+), with a large indirect contribution (82%) from PR status. This is consistent with the ER transcriptional control of PR[39] and a significantly higher mortality rate of ER+/PR- patients (11.8%) than ER+/PR+ patients (4.4%). Likewise, iMIIC infers a number of direct associations between the histology of primary tumors and other variables, Extended Data Fig. 8a, such as Age at diagnosis (in agreement with early reports[40]) and with synchro bilateral primaries (detected within 6 months of first diagnosis) which are almost twice more likely to occur when lobular carcinoma is present, Extended Data Fig. 8b. By contrast, no significant association is found with contralateral primary tumors detected more than 6 months after diagnosis, Extended Data Fig. 8c.

**Surgery and subsequent treatment subnetwork.** Interestingly, iMIIC also uncovers the central role of Surgery on the precise characterisation of primary tumors, Fig. 4a. For instance, iMIIC uncovers a somewhat unexpected genuine causal link from Surgery to Histology, which reflects that histological types are frequently refined after surgery by the pathologist based on the surgical specimen. This is consistent with a significant increase in histological types including specific tissues after surgery such as Infiltrating duct mixed with other types of carcinoma (+77% after surgery), Infiltrating duct and lobular carcinoma (+48%), Infiltrating duct

carcinoma, NOS (+7.6%), and a corresponding decrease in more generic histological types such as Lobular carcinoma, NOS (-11%), Carcinoma, NOS (-91%), and Adenocarcinoma, NOS (-95%). Similarly, iMIIC rightly infers that the staging variable, AJCC7thN, is usually based on the pathological report following surgery, while not performing surgery (due to the presence of distant metastases at diagnosis or the patient's old age) leads to much more frequent unspecified breast quadrant localisation for primary tumor, Fig. 4a, *i.e.* 30.4% "Breast NOS" when surgery is not recommended *versus* 11.1% when it is performed, Fig. 4b.

Likewise, iMIIC uncovers the central role of Surgery on the therapeutic decisions about subsequent treatments, such as breast reconstruction and radiotherapy, Fig. 4a. While breast reconstruction indeed requires breast surgery, Fig. 4c, iMIIC also correctly infers that the Type of Surgery at the Primary Site (lumpectomy or mastectomy) largely depends on the personal choice of early stage breast cancer patients between breast conservation and reconstruction alternatives, Fig. 4a,d. Similarly, iMIIC rightly infers that radiotherapy is a frequent "consequence" of breast surgery, Fig. 4a, *i.e.* 53% *versus* 4% radiotherapy if surgery is performed or not, Fig. 4e, especially after lumpectomy (75%) to limit the risk of relapse after breast conservation surgery.

**Socio-economic subnetwork.** The full breast cancer network on Fig. 2a includes four socio-economic variables pertaining to the county of residence of each patient: Median Family Income, Median Household Income, Cost of Living Index and the Rural-Urban population size of each county. These four socio-economic variables actually form a fully connected subgraph (*i.e.* a clique), indicating their strong interdependencies, and are directly connected to a number of other variables, Fig. 5a. Interestingly, Vital Status is only connected to this county variable clique through Median Household Income, which is consistent with earlier reports on the association between life expectancy and incomes[41]. By contrast, all other patient specific variables connected to the county clique (such as tumor grade, radiotherapy, breast reconstruction, insurance) have in fact at least one link with Cost of Living Index, highlighting the healthcare system integration into the global economy. In particular, there is a direct association between higher cost of living and more favorable breast cancer prognosis (*e.g.* fewer invasive components at diagnosis). This is presumably due to better preventive healthcare including easier access to breast cancer screening centers and more comprehensive insurance coverage. Yet, there are also strong disparities between counties, as manifested by the opposite associations of Insurance and Radiotherapy with Median Family Income *versus* Cost of Living Index, Fig. 5a. These intriguing findings can be traced back to Los Angeles (L.A.) county, amounting to about 10% of the whole dataset, which presents a lower than average median family income (29-38% percentile range) despite a higher than average cost of living index (58-67% percentile range), Fig. 5b. This must have led to an exacerbated financial burden for many of the 39,089 breast cancer patients diagnosed in L.A. county between 2010 and 2016. Although 18% of these patients benefited from medicaid insurance (as compared to 10% in the whole dataset), many had to opt for affordable but limited private insurance including significant co-payment policies or even to become uninsured especially before the application of the Affordable Care Act in January 2014 (3.4% uninsured in 2013 against 1.5% in 2014). As a result, many L.A. patients appear to have renounced to undergo expensive treatments. In particular, only 32.6% of patients underwent radiotherapy in L.A. as compared to 50% of patients nationwide excluding L.A. county, Fig. 5c, which can only be partly accounted for by county differences in under-reported radiotherapy of outpatients[30,31]. Moreover, an estimated 7% of L.A. patients even appear to have dropped out of therapy or moved to a different county not included in SEER database (against 1.5% nationwide, excluding L.A. county), based on the rapidly decreasing follow-up time distribution in L.A. as compared to the rest of the dataset, Fig. 5d. This corresponds to the fraction of patients having had their last medical contact less than a year after diagnosis and more than a year before the end of this study in December 2016.

## Discussion

Nationwide healthcare data, such as the SEER breast cancer dataset analyzed here, are especially interesting from a methodological point of view; they provide real-life benchmark datasets, which can help assess the reliability of causal discovery methods on real-life data, as most cause-effect predictions can be validated or dismissed, based on expert knowledge, clinical practice or possible data collection and selection biases. Besides, the interpretability of Machine Learning methods is particularly relevant for applications on clinical data, for which Artificial Intelligence assisted recommendations can hardly rely on black box classifiers only and need to be explainable in terms of intelligible rationales to medical practitioners[42]. Yet, beyond clinical

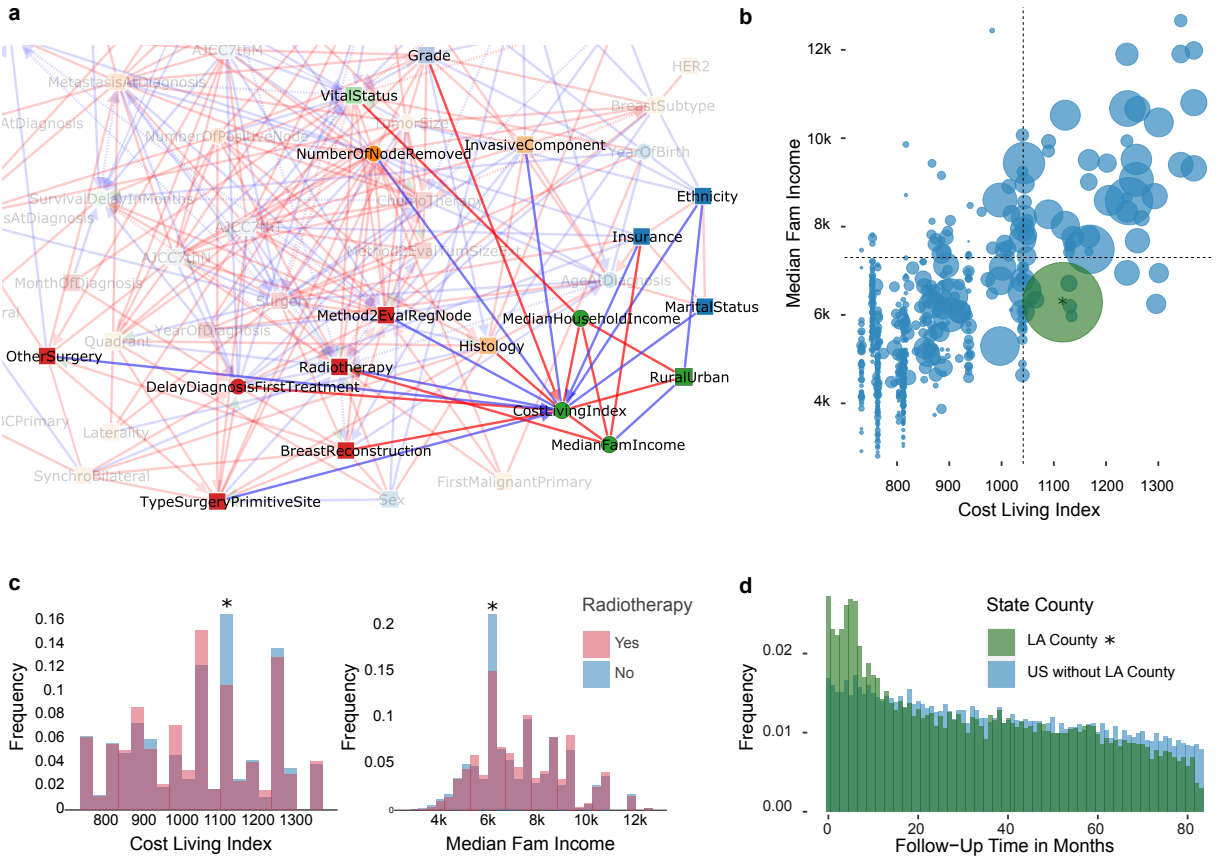



Figure 5: **Socio-economic subnetwork inferred by iMIIC from SEER breast cancer dataset**. **(a)** Subnetwork highlighting direct relations with socio-economic county variables (CostLivingIndex, MedianFamIncome, MedianHouseholdIncome, and RuralUrban). **(b)** Bubble plot of the joint distribution of Median Family Income and Cost of Living Index. The bubble area represents the number of patients in that county. Dashed lines correspond to the mean Cost of Living Index and mean Median Family Income. The green bubble with an asterisk corresponds to Los Angeles (L.A.) county which accounts for 10% of the full dataset. **(c)** Histograms of Cost of Living Index and Median Family Income grouped by Radiotherapy. Bins with an asterisk correspond to L.A. county. **(d)** Histograms of Follow-Up Time in Months for L.A. patients and for all other US counties included in SEER.

data, causal discovery methods have the potential to become essential Machine learning approaches to interpret diverse observational data in a wide range of domains, for which systematic perturbation experiments are not available due to practical, cost or ethical reasons. In particular, causal discovery can guide biological research by predicting the causal effects of specific interventions[43], such as gene expression or gene silencing, which can then be probed by targeted siRNA, gene knock-out or CRISPR-based editing experiments.

In the context of SEER's breast cancer dataset, iMIIC uncovers many expected causal relations, such as the adverse consequence of metastasis and the protecting effect of ER+ status on death due to breast cancer, or the fact that year of birth is the primary reason for death due to other causes by the end of the study. On the other hand, the effects of insurance coverage or marital status, which have been reported to reduce the risk of death due to breast cancer, are found to be entirely indirect and mainly mediated by treatments (60-80%), notably, surgery (>50%). In fact, surgery appears as the cornerstone of breast cancer therapy by first helping refine histological types, then guide therapeutic decisions on radiotherapy and breast reconstruction and ultimately prolong the survival delays of patients. Yet, iMIIC also correctly infers that the type of surgery (lumpectomy or mastectomy) at the primary site largely depends on the personal choice of early stage breast cancer patients between breast conservation or reconstruction alternatives. By contrast, other treatments, such as radiotherapy and chemotherapy, seem to have less decisive impacts on breast

cancer outcome, which might be due in part to some under-reported treatment information in the SEER database[30,31]. Radiotherapy even appears to be a consequence, not a cause, of vital status, suggesting that early death within the first few months after diagnosis may prevent radiotherapy for some patients who might have otherwise received this treatment, have they lived longer. Finally, iMIIC recovers direct associations between socio-economic county variables (such as median family income and cost of living index) and patient specific variables (such as tumor grade, radiotherapy, breast reconstruction, insurance), highlighting the healthcare system integration into the global economy. While higher costs of living are on average associated to more favorable cancer prognosis, presumably due to better preventive healthcare and more comprehensive insurance coverage, iMIIC also uncovers large disparities between family income and cost of living indices across counties (*e.g.* for L.A. county), leading to exacerbated financial burden with patients giving up expensive treatments or even dropping out of treatment.

In summary, iMIIC is a general causal discovery method, which uncovers direct and possibly causal relations as well as network consistent indirect effects for a broad range of biological and clinical data. Importantly, iMIIC is fully unsupervised and does not need preconceived hypothesis nor expert knowledge. In particular, iMIIC automatically adjusts for measured confounders (in the form of indirect contributions) and distinguishes genuine causes from putative and latent causal effects by either ruling out or highlighting the effect of unmeasured confounders for each causal edge (Box 1). While iMIIC is not immune to possible data collection and selection biases, which can affect observational data, it is based on a robust information theoretic framework, making it particularly reliable to interpret challenging types of data, such as heterogeneous data including combination of continuous and categorical variables integrated from different sources (*e.g.* clinical, personal, socio-economic data, as demonstrated here) or different experimental techniques (*e.g.* single cell transcriptomics[6,43,44] and imaging data[8]). In principle, iMIIC could be applied to a wide range of other domains to uncover causal relations and quantify indirect contributions when only observational data is available. With the advent of virtually unlimited datasets in many data science domains, scalable causal discovery methods are much needed and we believe that iMIIC can bring unique insights based on causal interpretation in many data science applications.

## Methods

**Overview and limitations of constraint-based methods.** Constraint-based methods proceed through successive steps, outlined in Fig. 1b, whose accuracy ultimately conditions the reliability and interpretability of the final causal graphical model. Starting from a fully connected graph, their first step consists in removing, iteratively, all dispensable edges whenever two variables are marginally independent or conditionally independent given a so-called separating set of conditioning variables. Positive (resp. negative) partial correlations are represented with red (resp. blue) edges in Fig. 1b and all other network figures. The rationale behind this first step is that all statistical associations between disconnected variables in the predicted graph should be graphically interpretable in terms of indirect paths through their separating set. This is, however, frequently not the case in practice[16].

The second step then consists in orienting some of the edges of the undirected graph (named skeleton) obtained at the first step, based on the signature of causality in observational data. This amounts to orient so-called "v-structures" as, $X \rightarrow Z \leftarrow Y$, whenever the edge $X - Y$ has been removed without including a common neighbor $Z$ of $X$ and $Y$ in their separating set, $\boldsymbol{S}$. The converging orientations of such a v-structure graphically indicate that $Z$ cannot be a cause of neither $X$ nor $Y$, which would otherwise require $Z$ to be included in the separating set, $\boldsymbol{S}$. However, this does not imply that $X$ (or $Y$) is an actual cause of $Z$, which also requires to rule out the possibility that the direct link between $X$ and $Z$ (or $Y$ and $Z$) might in fact originate from an unmeasured confounder, that is, from a latent common cause, $L$, unobserved in the dataset, *i.e.* $X \dashleftarrow L \dashrightarrow Z$, as described with an intuitive example in Box 1. Finally, the third step aims at propagating the orientations of v-structures to downstream edges, to fulfill the assumptions of the underlying graphical model class of constraint-based methods.

However, while traditional constraint-based methods have been shown to be theoretically sound and complete given an unlimited amount of data[5], they lack robustness on finite datasets, as their long series of uncertain decisions lead to an accumulation of errors, which limit the reliability of the final networks. In particular, spurious conditional independences, stemming from coincidental combinations of conditioning variables, lead to many false negative edges and, ultimately, limit the accuracy of inferred orientations.

**Overview and limitations of MIIC method.** The recent causal discovery method, MIIC, combines constraint-based and information-theoretic frameworks to learn more robust causal graphical models[6,8]. To limit the accumulation of errors in removing dispensable edges, MIIC does not directly attempt to uncover conditional independences but, instead, iteratively substracts the most significant information contributions of successive contributors, $A_1$, $A_2$, ..., $A_n$, from the mutual information between each pair of variables, $I(X;Y)$, as,

$$I(X;Y) - I(X;Y;A_1) - I(X;Y;A_2|A_1) - \cdots - I(X;Y;A_n|\{A_i\}_{n-1}) = I(X;Y|\{A_i\}_n) \quad (1)$$

where $I(X;Y;A_k|\{A_i\}_{k-1}) > 0$ is the *positive* information contribution from $A_k$ to $I(X;Y)$, that is not dependent on the first $k-1$ collected variables, $\{A_i\}_{k-1}$[44,45]. Conditional independence is eventually established when the residual conditional mutual information on the right hand side of Eq. 1, $I(X;Y|\{A_i\}_n)$, becomes smaller than a complexity term, *i.e.* $k_{X;Y|\{A_i\}}(N) \geqslant I(X;Y|\{A_i\}_n) \geqslant 0$, which dependents on the considered variables and sample size $N$. This complexity term also defines size corrected (or "regularized") conditional mutual information as,

$$I'(X;Y|\{A_i\}_n) = I(X;Y|\{A_i\}_n) - k_{X;Y|\{A_i\}}(N) \quad (2)$$

which become *negative* under conditional independence (*i.e.* $I'(X;Y|\{A_i\}_n) \leqslant 0$), that is, whenever sufficient and significant indirect positive contributions could be iteratively collected in Eq. 1 to warrant the removal of edge $XY$.

This leads to an undirected skeleton, which MIIC then (partially) orients based on the sign and amplitude of the regularized conditional 3-point information terms[6,45], corresponding to the difference between regularized conditional mutual information terms.

$$I'(X;Y;Z|\{A_i\}) = I'(X;Y|\{A_i\}) - I'(X;Y|\{A_i\},Z) \quad (3)$$

In particular, negative conditional 3-point information terms, $I'(X;Y;Z|\{A_i\}) < 0$, correspond to the signature of causality in observational data[45] and lead to the prediction of a v-structure, $X \to Z \leftarrow Y$, if $X$ and $Y$ are not connected in the skeleton (with $I'(X;Y|\{A_i\}) \leqslant 0$). By contrast, a positive conditional 3-point information term, $I'(X;Y;Z|\{A_i\}) > 0$, implies the absence of a v-structure and suggests to propagate the orientation of a previously directed edge $X \to Z - Y$ as $X \to Z \to Y$ (with $I'(X;Y|\{A_i\},Z) \leqslant 0$), to fulfill the assumptions of the underlying graphical model class.

In practice, MIIC's strategy to circumvent spurious conditional independences significantly improves the sensitivity or recall, that is, the fraction of correctly recovered edges, compared to traditional constraint-based methods, Extended Data Fig. 3. However, the original MIIC method still presents a number of limitations, such as a lower reliability in predicting edge orientation than edge presence, a limited scalability with continuous or mixed-type data, a remaining ambiguity on the "putative" *versus* "genuine" causal nature of oriented edges, and a possible inconsistency of separating sets with respect to indirect paths in the inferred network. The advanced iMIIC method overcomes all these limitations, as detailed in the remaining Methods' sections.

**Improved reliability of iMIIC inferred orientations.** While the original MIIC significantly outperforms traditional constraint-based methods in inferring reliable orientations, a substantial loss in precision usually remains between MIIC skeleton and oriented graph predictions, Extended Data Fig. 3. This is due to orientation errors originating from inconsistent v-structures, $X \to Z \leftarrow Y$, whose middle node $Z$ could also be included in the separating set of the unconnected pair $\{X,Y\}$, in contradiction with the head-to-head meeting of the v-structure. In particular, for discrete variables with (too) many levels, complexity terms can easily outweigh (conditional) mutual information for weakly dependent variables. As a result, original MIIC tends to infer some v-structure orientations, $X \to Z \leftarrow Y$, for which both (conditional) mutual information terms in Eq. 3 are negative, *i.e.* $I'(X;Y|\{A_i\}) < I'(X;Y|\{A_i\},Z) < 0$, suggesting that $Z$ could in fact be included in a separating set of the $\{X,Y\}$ pair, in contradiction with the inferred v-structure, $X \to Z \leftarrow Y$. To circumvent this issue, iMIIC implements more conservative orientation rules by essentially treating categorical and continuous variables alike, based on a general mutual information supremum principle[14,15], outlined below. In practice, iMIIC rectifies all negative regularized (conditional) mutual information, defining (conditional) independence (*e.g.* $I'(X;Y|\{A_i\}) \leqslant I'(X;Y|\{A_i\},Z) \leqslant 0$), to null values instead (*i.e.* $I'(X;Y|\{A_i\}) = I'(X;Y|\{A_i\},Z) = 0$), based on Theorem 1, below. This leads to vanishing conditional 2-point and 3-point information in Eq. 3, for the example above, and prevents the orientation of the inconsistent v-structure.

**General mutual information supremum principle.** Estimating (conditional) mutual information between continuous or mixed-type variables is notoriously more challenging than between categorical variables[46,47]. Original MIIC computes regularized mutual information between continuous or mixed-type variables through an optimum discretization scheme, based on a general mutual information supremum principle[14] regularized for finite datasets and using an efficient $\mathcal{O}(N^2)$ dynamic programming algorithm[8]. This approach finds optimum partitions, $\mathcal{P}$ and $\mathcal{Q}$, specifying the number and positions of cut-points of each continuous variable, $X$ and $Y$, to maximize the regularized mutual information between them,

$$I'(X;Y) = \sup_{\mathcal{P},\mathcal{Q}} I'([X]_{\mathcal{P}};[Y]_{\mathcal{Q}}) \quad (4)$$

Such optimization-based estimates of mutual information are at par with alternative distance-based k-nearest neighbor (kNN) approaches[46,47] but have also the unique advantage of providing an effective independence test to identify independent continuous or mixed-type variables[8]. This is achieved when partitioning $X$ and $Y$ into single bins maximizes the regularized mutual information in Eq. 4, which vanishes exactly in this case, *i.e.* $I'(X;Y) = I'([X]_1;[Y]_1) = 0$ if $I'([X]_\mathcal{P};[Y]_\mathcal{Q}) \leqslant 0$ for all partitions $\mathcal{P}, \mathcal{Q}$. By contrast, kNN estimates still need an actual independence test to decide whether some variables are effectively independent or not, as kNN mutual information estimates are never exactly null.

Yet, the optimum partitioning principle (Eq. 4) only applies to mutual information[14], *not* conditional mutual information, which need to be estimated through the *difference* between optimum regularized mutual information terms, as $I'(X;Y|U) = I'(Y;\{X,U\}) - I'(Y;U) = I'(X;\{Y,U\}) - I'(X;U)$[8]. As a result of numerical approximation, the regularized conditional mutual information estimates between conditionally independent variables can sometime be negative and lead to inconsistent v-structure orientations, as discussed for discrete data above.

The general mutual information supremum principle[14], regularized for finite datasets in Eq. 4, is theoretically valid for any type of variable, not just continuous variables. In particular, it could be applied to datasets including discrete or categorical variables with (too) many levels. This would result in the merging of rare levels to better estimate mutual information and conditional mutual information between weakly dependent discrete variables. Ultimately, mutual information estimates between independent discrete variables should lead to the merging of each variable into a single bin, thereby, resulting in regularized mutual information estimates to vanish exactly in this case, as observed for continuous variables. As a result, optimum regularized mutual information should be non-negative as well as, by extension, regularized conditional mutual information, as proved below.

***Theorem 1.***[15] Regularized (conditional) mutual information derived from the general mutual information supremum principle are non-negative.

*Proof.* We first address optimum regularized mutual information, noting that $I'(X;Y) \geqslant I'([X]_1;[Y]_1) = 0$, where $[X]_1$ and $[Y]_1$ are the $X$ and $Y$ variables partitioned into single bins, which leads to a vanishing regularized mutual information, as both mutual information and complexity cost are null for single bin partitions[45]. Then, regularized conditional mutual information is defined as the *difference* between optimum regularized mutual information terms as, $I'(X;Y|U) = I'(Y;\{X,U\}) - I'(Y;U) = I'(X;\{Y,U\}) - I'(X;U)$. However, partitioning $X$ and $Y$ into a single bin leads to $I'(Y;\{X,U\}) \geqslant I'(Y;\{[X]_1,U\}) = I'(Y;U)$ and $I'(X;\{Y,U\}) \geqslant I'(X;\{[Y]_1,U\}) = I'(X;U)$ thus implying $I'(X;Y|U) \geqslant 0$ □

Based on Theorem 1, iMIIC rectifies negative values of regularized (conditional) mutual information, indicating (conditional) independence, to null values instead. This simple modification is found to significantly enhance the reliability of iMIIC predicted orientations, in particular for datasets with high proportions of discrete variables, with only a small sensitivity loss compared to MIIC original orientation rules, Fig. 1c.

**Scalable computations of multivariate information and iMIIC orientation scores.** The running time of the original MIIC algorithm scales linearly with sample size for discrete datasets[6] but at best quadratically with sample size for continuous or mixed-type datasets[8], due to a $\mathcal{O}(N^2)$ dynamic programming optimization of the number and positions of cut points to estimate (conditional) multivariate information. This quadratic scaling becomes prohibitive for very large datasets, such as the SEER dataset analyzed here, which contains nearly 400,000 breast cancer patients. To circumvent this scalability issue, iMIIC enforces a maximum number of 50 bins, so that the overall optimisation of multivariate information estimates remains close to linear in terms of sample size, see Extended Data Figs. 4-6.

A second scalability issue concerns the estimation of orientation probabilities by the original MIIC, which are numerically too close to be reliably compared for very large datasets and require to introduce scalable orientation scores and novel definitions of induced tail and head orientation scores, as detailed now.

*V-structure orientation scores.* Head orientation probabilities of v-structures, $X *\!\!\rightarrow Z \leftarrow\!\!* Y$, are computed from negative regularized (conditional) 3-point information, $I'(X;Y;Z \mid \{A_i\}) < 0$, as,[6]

$$P(x *\!\!\rightarrow \underline{z}) = P(\underline{z} \leftarrow\!\!* y) = \frac{1 + e^{NI'(X;Y;Z|\{A_i\})}}{1 + 3e^{NI'(X;Y;Z|\{A_i\})}} \geqslant \frac{1}{2} \tag{5}$$

where the end mark $(*)$ stands either for a head $(>)$, a tail $(-)$ or is undefined $(\circ)$, and $e^{NI'(X;Y;Z|\{A_i\})}$ corresponds to the probability ratio between a non-v-structure and a v-structure, $e^{NI'(X;Y;Z|\{A_i\})} = P_{\rightarrow -}/P_{\rightarrow\leftarrow} = P_{-\leftarrow}/P_{\rightarrow\leftarrow} = P_{--}/P_{\rightarrow\leftarrow}$. However, due to numerical precision Eq. 5 cannot rank orientation probabilities that are too close to 1 for large $N$ and iMIIC resorts instead to equivalent v-structure orientation scores,

$$\begin{aligned} \text{score}_\text{v} &= -NI'(X;Y;Z \mid \{A_i\}) + \log 1\text{p}\big(e^{NI'(X;Y;Z|\{A_i\})}\big) - \log 2 \\ P(x *\!\!\rightarrow \underline{z}) &= P(\underline{z} \leftarrow\!\!* y) = \frac{1}{1 + e^{-\text{score}_\text{v}}} \end{aligned} \tag{6}$$

which enable the ordering of orientation probabilities, $P_1$ and $P_2$ between alternative v-structures ($v_1$ and $v_2$), even for very large $N$, as $0 \leqslant \text{score}_1 < \text{score}_2 < \infty$ is equivalent to $0.5 \leqslant P_1 < P_2 < 1$.

*Induced tail and head orientation scores.* Similarly, induced orientation probabilities originating from an existing arrowhead $\underline{z} \leftarrow\!* \, y$ can be estimated through the following probability decomposition formula[6],

$$P(x *\!\!-\!\!\bullet \underline{z}) = P(x *\!\!-\!\!\bullet \underline{z} | \underline{z} \leftarrow\!* \, y) P(\underline{z} \leftarrow\!* \, y) + P(x *\!\!-\!\!\bullet \underline{z} | \underline{z} -\!* \, y) P(\underline{z} -\!* \, y) \tag{7}$$

where $\bullet$ stands for a tail [resp. a head] depending on the positivity [resp. negativity] of $I'(X;Y;Z \mid \{A_i\})$ and a corresponding (conditional) independence $I'(X;Y \mid \{A_i\}, Z) \leqslant 0$ [resp. $I'(X;Y \mid \{A_i\}) \leqslant 0$].

However, using the full probability decomposition above can lead to a higher confidence in tail or head induced probabilities than in the head probabilities they derive from, due to the Markov equivalence of non-v-structures. In addition, induced tail / head probabilities become numerically difficult to compare for large $N$, as Eq. 7 cannot be expressed in the form of Eq. 6. To circumvent these issues and capture the rationale that the confidence in induced tail / head orientations can only be lower than the confidence in the arrowhead from which they derive, iMIIC redefines the induced tail / head probabilities by retaining only the first term in the probability decomposition above, that is, by assuming that the arrowhead $\underline{z} \leftarrow\!* \, y$ exists,

$$\begin{aligned} P(x *\!\!-\!\!\bullet \underline{z}) &= P(x *\!\!-\!\!\bullet \underline{z} | \underline{z} \leftarrow\!* \, y) P(\underline{z} \leftarrow\!* \, y) \\ &= \frac{1}{1 + e^{-N|I'(X;Y;Z|\{A_i\})|}} \times \frac{1}{1 + e^{-\text{score}_\text{v}}} = \frac{1}{1 + e^{-\text{score}_\text{i}}} \end{aligned} \tag{8}$$

where we introduced a rectified induced $\text{score}_\text{i}$,

$$\begin{aligned} \text{score}_\text{i} &= \max\left(0, \text{m} - \text{log1p}\left(e^{-\text{M}+\text{m}} + e^{-\text{M}}\right)\right) \\ \text{m} &= \min\left(N|I'(X;Y;Z \mid \{A_i\})|, \text{score}_\text{v}\right) \\ \text{M} &= \max\left(N|I'(X;Y;Z \mid \{A_i\})|, \text{score}_\text{v}\right) \end{aligned} \tag{9}$$

to enable a global numerical ranking of v-structure orientation and induced orientation probabilities even for very large $N$ with $0.5 \leqslant P_1 < P_2 < 1$ corresponding to $0 \leqslant \text{score}_1 < \text{score}_2 < \infty$.

In addition, when orientation propagation is enforced (*i.e.* step 3 in Fig.1b), an induced tail probability can also be "propagated", as a head probability, to the other end of the edge, if its end mark is still undefined, *i.e.*, $P(\underline{x} \leftarrow z) = P(x \circ\!\!- \underline{z})$. However, this orientation propagation rule does not rely on specific information in the available data but rather aims at fulfilling the structural assumptions of benchmark graphical models. Hence, propagation has been applied in benchmark comparisons (Fig. 1c,d, Extended Data Figs. 3-6) but discarded to analyze real-life data (Figs. 2-5, Extended Data Figs. 7,8), in order to ensure that causal discovery on real-life applications is solely based on information actually contained in the available data.

**Orientation confidence and causal nature of edges.** Having fully ordered orientation probabilities, even for very large $N$, enables to implement edge orientations in decreasing order of confidence rather than any arbitrary order, as implemented in traditional constraint-based methods. In addition, iMIIC allows also to use an orientation confidence threshold $1 > \beta \geqslant 0.5$ to enhance the precision of predicted head and tail orientations and, thereby, our confidence in the causal nature of oriented edges. Hence, a genuine causal relation (represented with a green arrow-head) is predicted if the edge can be assigned both significant head and tail probabilities, $P_h > \beta$ and $P_t > \beta$, while a putative causal relation is inferred if only one significant head probability can be assessed given the available observational data, *i.e.* $P_h > \beta$ and $P_t \leqslant \beta$. Similarly, a bidirected edge, suggesting the effect of an unobserved common cause, is predicted for two significant head probabilities, while all other cases are graphically represented as undirected edges. In practice, orientation precision threshold $\beta$ mostly impacts the orientations derived from small datasets and has little effects on large datasets such as SEER presented here. All causal discovery benchmark results have also been obtained without enhancing orientation precision (*i.e.* using $\beta = 0.5$) which yields a better balance between precision and recall for all sample size. Finally, iMIIC also allows to include prior knowledge about certain head or tail orientations in graphical models, for instance, to specify contextual variables (*e.g.* sex, year of birth), which cannot be the consequence of other observed or non-observed variables in the dataset, as outlined in the main text.

**Indirect path consistency and information contribution.** As mentioned in the overview and limitations, above, traditional constraint-based methods, as well as, the original MIIC method do not control for the global structural consistency of their inferred networks. In particular, there is no guarantee that the separating sets identified during the iterative removal of edges (step 1) remain consistent in terms of indirect paths in the final network. To this end, iMIIC adapts a novel algorithmic scheme[16] to ensure that all separating sets identified to remove dispensable

edges are consistent with the final inferred graph. It is achieved by repeating the constraint-based structure learning scheme, iteratively, while searching for separating sets that are consistent with the graph obtained at the previous iteration, as outlined in Fig. 1b. We define two levels of indirect path consistency: skeleton *versus* orientation consistencies. Skeleton consistency guarantees that any node in a separating set is on an indirect path between the extremities of the corresponding removed edge (regardless of orientations along the path), while orientation consistency further enforces that each node in a separating set is a non-descendent neighbor of at least one of these extremities. Importantly, implementing skeleton or orientation consistency of separating sets can be done at a limited complexity cost, through the use of block-cut tree decomposition of graphs[16]. All in all, iMIIC indirect path consistency improves the interpretability of the inferred network in terms of indirect effects, which are also quantified with indirect information contributions, based on Eq. 1 including finite size corrections from Eq. 2,

$$\mathrm{IndC}(A_k; XY|\{A_i\}_{k-1}) = \frac{I'(X;Y;A_k|\{A_i\}_{k-1})}{I'(X;Y)} \tag{10}$$

with $\sum_k^n \mathrm{IndC}(A_k; XY|\{A_i\}_{k-1}) = 100\% - I'(X;Y|\{A_i\}_n)/I'(X;Y)$, where $I'(X;Y|\{A_i\}_n)/I'(X;Y)$ is the residual fraction of mutual information (*i.e.* not accounted for by $A_1, A_2, \cdots, A_n$ indirect contributions given by Eq. 10), which vanishes if the $XY$ edge has been removed, that is, if $I'(X;Y|\{A_i\}_n) = 0$, after negative value rectification.

**Data generation and benchmarks.** Synthetic datasets were simulated using a network structure inferred from a 10,000 patient subset of the full SEER dataset of breast cancer patients, leading to a network average connectivity of 5. Random network skeletons sharing the same SEER-like skeleton degree distribution were first obtained using a Monte Carlo graph generation algorithm[48]. These skeletons were subsequently oriented to obtain Directed Acyclic Graphs using a random ordering of their nodes and assigning various proportions of discrete *versus* continuous variables. The marginal distributions of variables without parents were chosen to resemble typical SEER-like marginal distributions, Extended Data Fig. 1, and the other variables were simulated using mixed-type structural equation models (SEMs)[8], see *e.g.* Extended Data Fig. 2. For each discrete node proportion (decile steps), 25 networks were generated with 50,000 simulated samples each.

For evaluation purposes, network reconstruction was treated as a binary classification task and classical performance measures, Precision, Recall and F-score, were computed to evaluate (i) skeleton, (ii) CPDAG and (iii) oriented-edge subgraph reconstructions. CPDAG scores use the same metrics as skeleton scores but rating as "false positive" the erroneous orientation of non-oriented edges and the non-orientation or opposite orientation of oriented edges in the CPDAG. However, these errors are not equivalent from a causal discovery perspective. By contrast, oriented-edge subgraph scores, highlighted in the benchmark comparisons, are restricted to the subgraphs containing only oriented edges in the theoretical CPDAG *versus* the inferred graph. These oriented-edge scores are designed to specifically assess the method performance on causal discovery, that is, on the oriented edges which can in principle be learnt from observational data *versus* those effectively predicted by the causal structure learning method. Five causal discovery methods able to analyze mixed-type datasets have been compared over SEER-like benchmarks:

- *Interpretable MIIC (iMIIC)* was run with default parameters for all settings.
- *Original MIIC*[7,8] was run with default parameters for all settings (Fig. 1c and Extended Data Fig. 3).
- *PC*[17] from the *pcalg* package[18] was run with the stable option[49] and either majority rule[49] (Extended Data Fig. 3) or conservative rule[19] (Fig. 1d and Extended Data Fig. 4) for orientations. The "ci.test" function from the *bnlearn* package[50] was used as independence test for mixed-type data (with either "mi-cg" option for discrete against continuous variables, "mi" for discrete against discrete variables or "mi-g" for continuous against continuous variables) and the threshold for significance testing was set to the default $\alpha = 0.01$.
- *causalMGM*[20] was run with the *rCausalMGM* R package. The initial graph was computed using the mgm() function with each of the 3 lambda parameters equal to 0.05 and the orientations were then obtained with the pcMax() function with default $\alpha = 0.01$ parameter.
- *MXM*[21], a mixed-PC constraint-based method, was run using the *MXM* R package. The graph was obtained using the pc.skel() function for skeleton with the "comb.mm" independence test and the default $\alpha = 0.01$ threshold for significance testing and with the pc.or() function for orientations.

*Computation time.* Benchmarks were stopped when the average computation time of a method reached 1 hour per network with high proportion of continuous variables (resp. about 10 minutes per network with low proportion of continuous variables), corresponding to a maximum running time of about 115h for the 250 generated networks at each sample size.

*Benchmark results.* The performance of iMIIC has been benchmarked against state-of-the-art constraint-based methods: PC, causalMGM and MXM, on SEER-like benchmark datasets with different proportions of discrete variables,

Fig. 1d and Extended Data Figs. 4-6. Results for datasets with 80% discrete variables, corresponding to the actual proportion in the real-life SEER breast cancer dataset, are discussed in the main text. Similarly, for larger proportions of continuous variables, Fig. 1d and Extended Data Figs. 4-6 demonstrate that iMIIC greatly outperforms the reliability and sensitivity of predicted orientations against state-of-the-art constraint-based methods. For instance, for SEER-like benchmark datasets with only 20% of discrete variables, iMIIC already reaches 81% (resp. 64%) in precision (resp. F-score), for $N = 10^3$, against 53% (29%) for conservative PC, 50% (40%) for causalMGM and 29% (25%) for MXM. For $N = 10^4$, iMIIC reaches 88% (78%) in precision (F-score), against about 60% (45%) for conservative PC, 52% (50%) for causalMGM and 22% (28%) for MXM. Finally, iMIIC reaches 86% (81%) for $N = 10^5$, which is beyond the sample size attainable by other methods.

## Data availability

The dataset of breast cancer patients was obtained from the Surveillance, Epidemiology and End Results programme, which can be accessed at `https://seer.cancer.gov/seertrack/data/request/`.

## Code availability

Causal discovery using iMIIC was performed on the open access server `https://miic.curie.fr` or running the R package available at `https://github.com/miicTeam/miic_R_package`. Other R packages used for benchmark comparisons are available at `https://r-forge.r-project.org/projects/pcalg`, `https://cran.r-project.org/web/packages/bnlearn`, `https://github.com/tyler-lovelace1/rCausalMGM` and `https://cran.r-project.org/web/packages/MXM`.

# References


1. Spirtes, P., Glymour, C. N., Scheines, R. & Heckerman, D. *Causation, prediction, and search* (MIT press, 2000).
2. Pearl, J. *Causality* (Cambridge university press, 2009).
3. Heckerman, D., Geiger, D. & Chickering, D. M. Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. *Mach. Learn.* **20,** 197–243 (1995).
4. Friedman, J., Hastie, T. & Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9,** 432–441 (2008).
5. Zhang, J. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artif. Intell.* **172,** 1873–1896 (2008).
6. Verny, L., Sella, N., Affeldt, S., Singh, P. P. & Isambert, H. Learning causal networks with latent variables from multivariate information in genomic data. *PLoS Comput. Biol.* **13,** e1005662 (2017).
7. Sella, N., Verny, L., Uguzzoni, G., Affeldt, S. & Isambert, H. MIIC online: a web server to reconstruct causal or non-causal networks from non-perturbative data. *Bioinformatics* **34,** 2311–2313 (2018).
8. Cabeli, V. *et al.* Learning clinical networks from medical records based on information estimates in mixed-type data. *PLOS Computational Biology* **16,** e1007866 (2020).
9. Howlader, N. *et al.* in *SEER Cancer Statistics Review* 1975–2016 (National Cancer Institute, 2018).
10. Peters, J., Mooij, J. M., Janzing, D. & Schölkopf, B. Causal Discovery with Continuous Additive Noise Models. *Journal of Machine Learning Research* **15,** 2009–2053 (2014).
11. Sackett, D. L. Bias in analytic research. *Journal of Chronic Diseases* **32,** 51–63 (1979).
12. Hernán, M. A., Hernández-Díaz, S. & Robins, J. M. A Structural Approach to Selection Bias. *Epidemiology* **15,** 615–625 (2004).
13. Prosperi, M. *et al.* Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence* **2,** 369–375 (2020).
14. Cover, T. M. & Thomas, J. A. *Elements of Information Theory* 2nd (Wiley, 2006).
15. Cabeli, V., Li, H., da Câmara Ribeiro-Dantas, M., Simon, F. & Isambert, H. *Reliable causal discovery based on mutual information supremum principle for finite datasets* in *why21 at 35rd Conference on Neural Information Processing Systems* (NeurIPS, 2021).

16. Li, H., Cabeli, V., Sella, N. & Isambert, H. *Constraint-based causal structure learning with consistent separating sets* in *33rd Conference on Neural Information Processing Systems* (NeurIPS, 2019).

17. Spirtes, P. & Glymour, C. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review* **9,** 62–72 (1991).

18. Kalisch, M., Mächler, M., Colombo, D., Maathuis, M. H. & Bühlmann, P. Causal inference using graphical models with the R package pcalg. *J. Stat. Softw.* **47,** 1–26 (2012).

19. Ramsey, J., Spirtes, P. & Zhang, J. *Adjacency-Faithfulness and Conservative Causal Inference* in *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence* (AUAI Press, 2006), 401–408.

20. Sedgewick, A. J. *et al.* Mixed graphical models for integrative causal analysis with application to chronic lung disease diagnosis and prognosis. *Bioinformatics* **35,** 1204–1212 (2018).

21. Tsagris, M., Borboudakis, G., Lagani, V. & Tsamardinos, I. Constraint-based causal discovery with mixed data. *International Journal of Data Science and Analytics* **6,** 19–30 (2018).

22. Harbeck, N. *et al.* Breast cancer. *Nature Reviews Disease Primers* **5,** 6 (2019).

23. Alaa, A. M., Gurdasani, D., Harris, A. L., Rashbass, J. & van der Schaar, M. Machine learning to guide the use of adjuvant therapies for breast cancer. *Nature Machine Intelligence* **3,** 716–726 (2021).

24. Lee, C. *et al.* Application of a novel machine learning framework for predicting non-metastatic prostate cancer-specific mortality in men using the Surveillance, Epidemiology, and End Results (SEER) database. *The Lancet Digital Health* **3,** e158–e165 (2021).

25. Welch, H. G., Prorok, P. C., O'Malley, A. J. & Kramer, B. S. Breast-Cancer Tumor Size, Overdiagnosis, and Mammography Screening Effectiveness. *New England Journal of Medicine* **375,** 1438–1447 (2016).

26. Leapman, M. S. *et al.* Mediators of Racial Disparity in the Use of Prostate Magnetic Resonance Imaging Among Patients With Prostate Cancer. *JAMA Oncology, Published online March 03* (2022).

27. Petito, L. C. *et al.* Estimates of Overall Survival in Patients With Cancer Receiving Different Treatment Regimens. *JAMA Network Open* **3,** e200452 (2020).

28. Nethery, R. C., Yang, Y., Brown, A. J. & Dominici, F. A causal inference framework for cancer cluster investigations using publicly available data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **183,** 1253–1272 (2020).

29. Wang, L. Mining causal relationships among clinical variables for cancer diagnosis based on Bayesian analysis. *BioData Mining* **8,** 13 (2015).

30. Park, H. S., Lloyd, S., Decker, R. H., Wilson, L. D. & Yu, J. B. Limitations and Biases of the Surveillance, Epidemiology, and End Results Database. *Current Problems in Cancer* **36,** 216–224 (2012).

31. Jagsi, R. *et al.* Underascertainment of radiotherapy receipt in Surveillance, Epidemiology, and End Results registry data. *Cancer* **118,** 333–341 (2011).

32. Chen, S.-Y. *et al.* Timing of Chemotherapy and Radiotherapy Following Breast-Conserving Surgery for Early-Stage Breast Cancer: A Retrospective Analysis. *Frontiers in Oncology* **10,** 571390 (2020).

33. Anderson, J. R., Cain, K. C. & Gelber, R. D. Analysis of survival by tumor response. *Journal of Clinical Oncology* **1,** 710–719 (1983).

34. Hernán, M. A. & Robins, J. M. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available: Table 1. *American Journal of Epidemiology* **183,** 758–764 (2016).

35. Han, X., Yabroff, K. R., Ward, E., Brawley, O. W. & Jemal, A. Comparison of Insurance Status and Diagnosis Stage Among Patients With Newly Diagnosed Cancer Before vs After Implementation of the Patient Protection and Affordable Care Act. *JAMA Oncology* **4,** 1713 (2018).

36. Ermer, T. *et al.* Understanding the Implications of Medicaid Expansion for Cancer Care in the US. *JAMA Oncology* **8,** 139 (2022).

37. Hinyard, L., Wirth, L. S., Clancy, J. M. & Schwartz, T. The effect of marital status on breast cancer-related outcomes in women under 65: A SEER database analysis. *The Breast* **32,** 13–17 (2017).

38. Zhai, Z. *et al.* Effects of marital status on breast cancer survival by age, race, and hormone receptor status: A population-based Study. *Cancer medicine* **8,** 4906–4917 (2019).

39. Bonéy-Montoya, J., Ziegler, Y. S., Curtis, C. D., Montoya, J. A. & Nardulli, A. M. Long-range transcriptional control of progesterone receptor gene expression. *Mol Endocrinol* **24,** 346–358 (2010).

40. Fisher, C. *et al.* Histopathology of breast cancer in relation to age. *British journal of cancer* **75,** 593–596 (1997).

41. Chetty, R. *et al.* The Association Between Income and Life Expectancy in the United States, 2001-2014. *JAMA* **315,** 1750 (2016).

42. Binder, A. *et al.* Morphological and molecular breast cancer profiling through explainable machine learning. *Nature Machine Intelligence* **3,** 355–366 (Mar. 2021).

43. Desterke, C. *et al.* Inferring Gene Networks in Bone Marrow Hematopoietic Stem Cell-Supporting Stromal Niche Populations. *iScience* **23,** 101222 (2020).

44. Affeldt, S., Verny, L. & Isambert, H. 3off2: A network reconstruction algorithm based on 2-point and 3-point information statistics. *BMC Bioinformatics* **17,** 12 (2016).

45. Affeldt, S. & Isambert, H. *Robust reconstruction of causal graphical models based on conditional 2-point and 3-point information* in *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence* (UAI, 2015), 42–51.

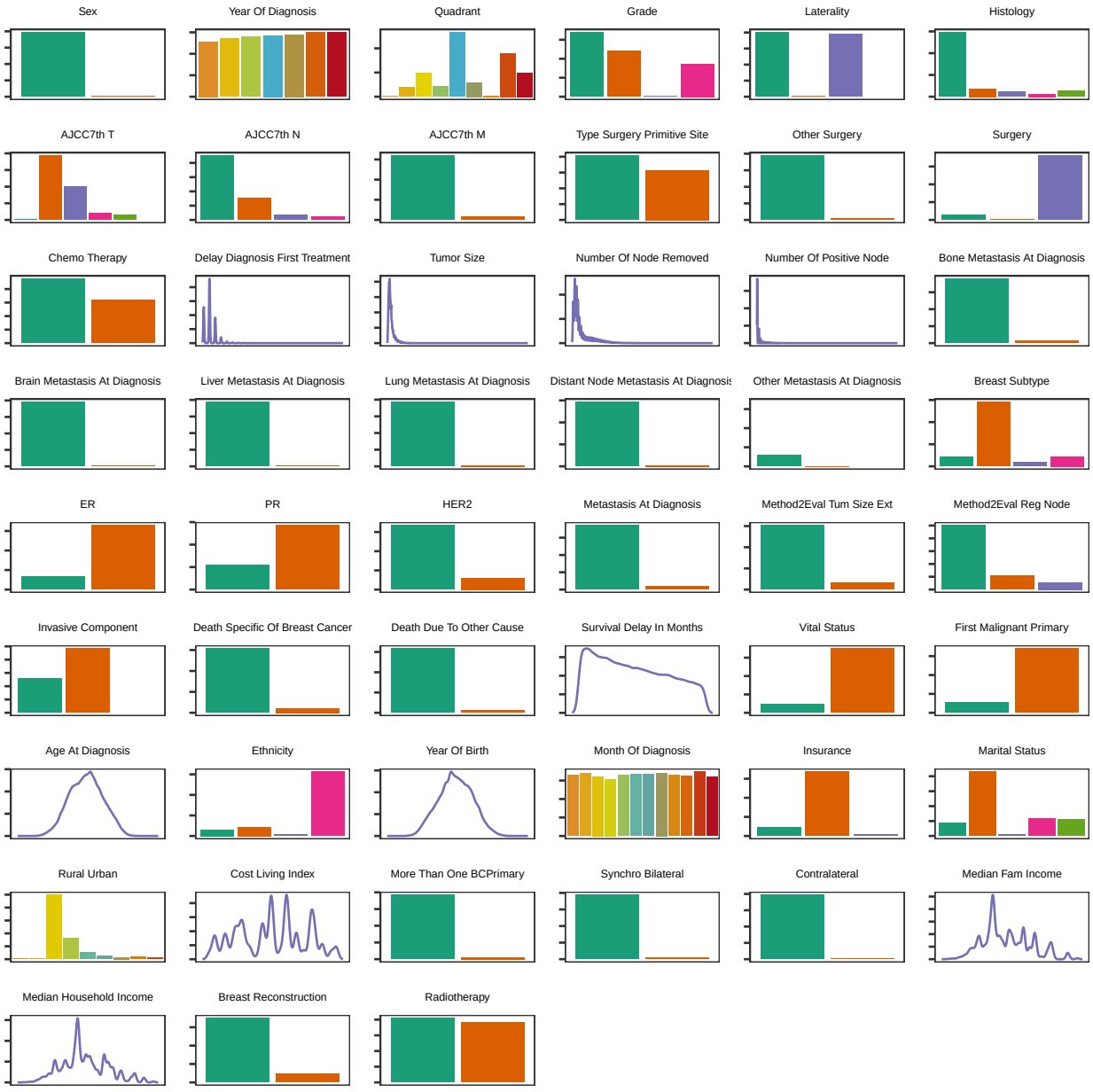
46. Kraskov, A., Stögbauer, H. & Grassberger, P. Estimating mutual information. *Phys. Rev. E* **69,** 066138 (2004).

47. Frenzel, S. & Pompe, B. Partial Mutual Information for Coupling Analysis of Multivariate Time Series. *Phys. Rev. Lett.* **99,** 204101 (2007).

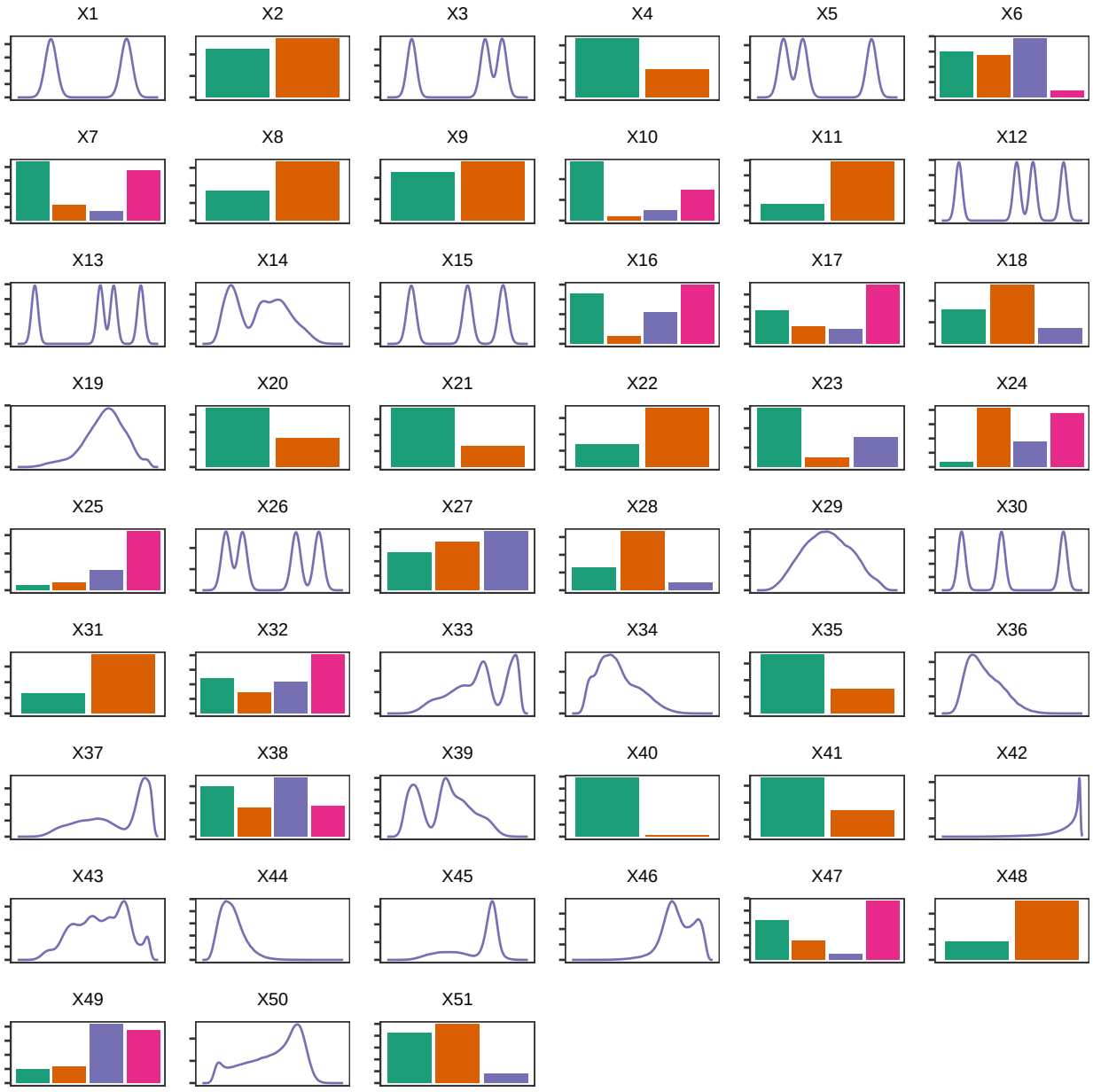
48. Viger, F. & Latapy, M. in *Lecture Notes in Computer Science* 440–449 (Springer Berlin Heidelberg, 2005).

49. Colombo, D. & Maathuis, M. H. Order-Independent Constraint-Based Causal Structure Learning. *J. Mach. Learn. Res.* **15,** 3741–3782 (2014).

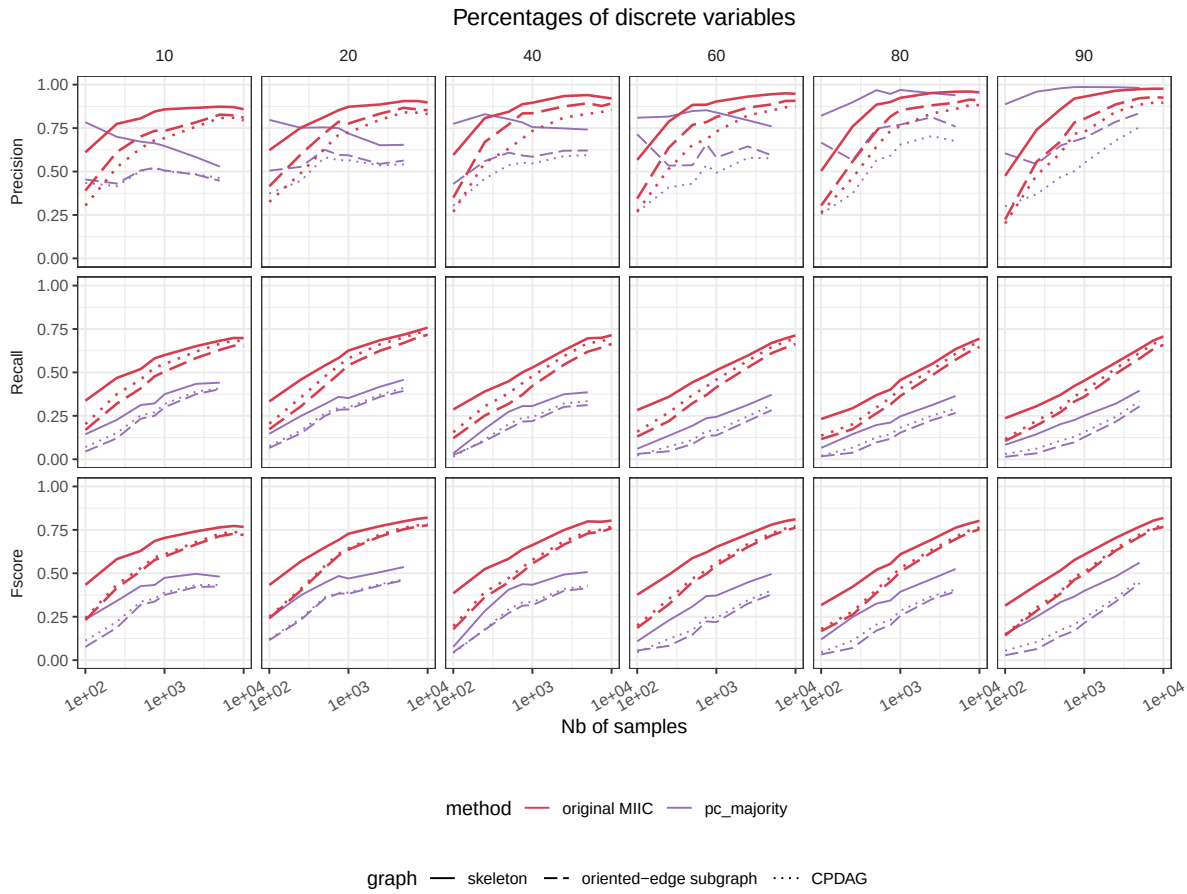
50. Scutari, M. Learning Bayesian Networks with the bnlearn R Package. *Journal of Statistical Software* **35,** 1–22 (2010).



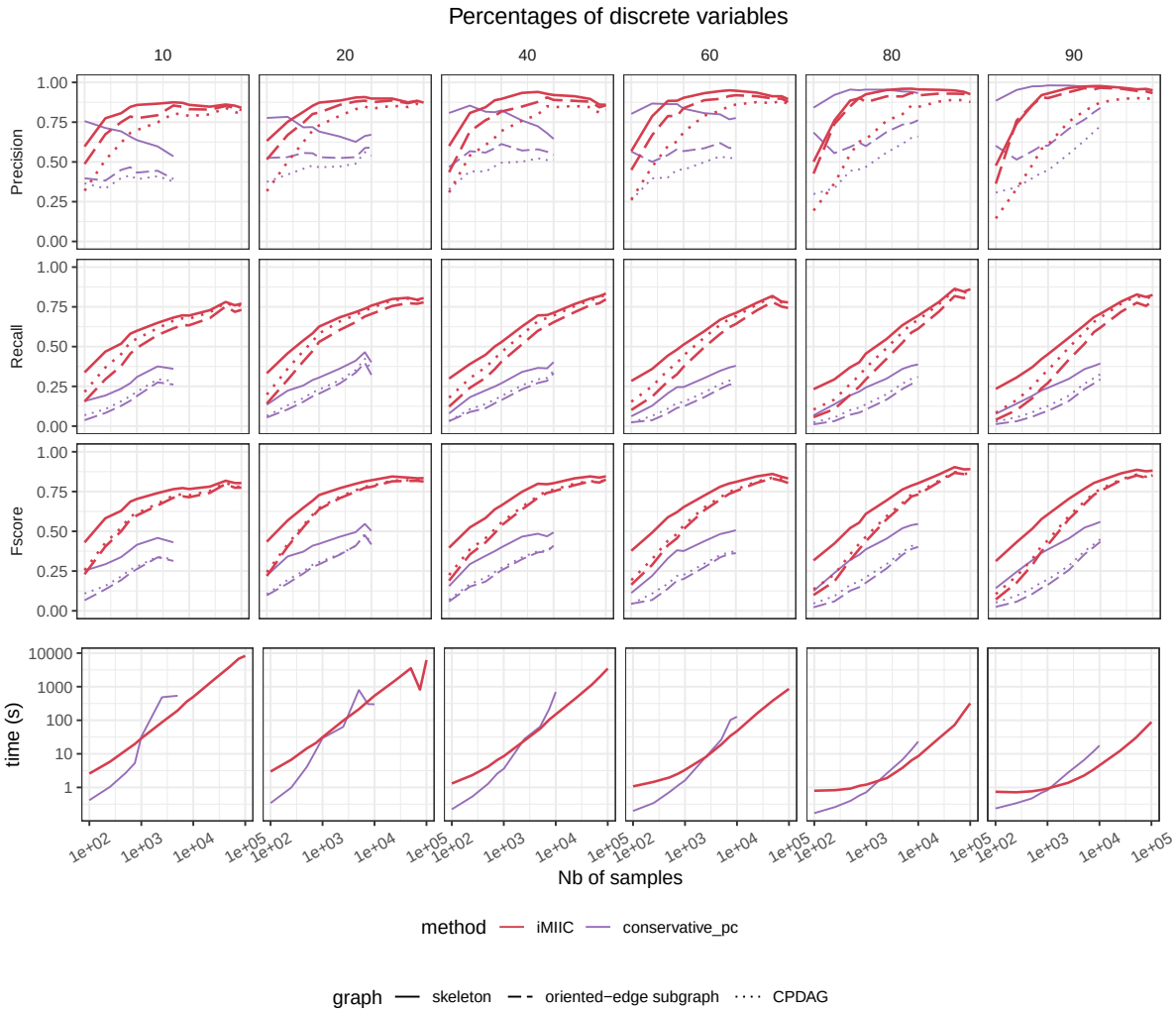

Extended Data Fig. 1: **Distributions of the 51 SEER variables selected for breast cancer**. There are 407,791 breast cancer records in SEER for the period 2010-2016, but only 396,179 distinct patients due to multiple breast primary tumors for some patients. For each patient, we selected the first breast primary tumor recorded in SEER and indicated the total number of breast cancer primaries during the 2010-2016 period in the variable *MoreThanOneBCPrimary*. *SynchroBilateral* was also engineered to identify patients who had tumors in both breasts diagnosed within less than 180 days of each other, while *Contralateral* identifies patients who had a subsequent tumor in the other breast diagnosed more than 180 days after the first breast tumor primary. Some categorical variables had some of their categories merged, either because these categories had the same general meaning or because they were too rare amongst patients (*i.e.* <0.1% of patients excluding those with missing data for the considered variable). These variables include *Ethnicity, TypeSurgeryPrimitiveSite, Surgery, OtherSurgery, OtherMetastasisAtDiagnosis, Insurance and Histology*. Hence, categories recorded in less than 0.1% of patients were merged and renamed to 'Other'. *BreastReconstruction* was engineered based on *TypeSurgeryPrimitiveSite* (*i.e.* SEER surgery code ranges 43-49, 53-59, 63-69, and 73-75 were assigned 'Yes', while other surgery codes were assigned 'No'). *Radiotherapy* was created from *Radiation sequence with surgery*, that has much fewer missing data (0.05%) than the original *Radiation* variable (49%). *TumorSize* merges two distinct variables that contained tumor sizes for years 2004-2015 and 2016+, respectively. Likewise, the largely missing 2016 information for the *MetastasisAtDiagnosis* variable was recovered based on information contained in specific metastasis variables (*i.e. BoneMetastasisAtDiagnosis, LungMetastasisAtDiagnosis, LiverMetastasisAtDiagnosis, BrainMetastasisAtDiagnosis, OtherMetastasisAtDiagnosis*). Finally, *MedianFamIncome* and *MedianHouseHoldIncome* are the average of these continuous variables over the periods 2007-2011, 2008-2012, 2009-2013, 2010-2014, 2011-2015, and 2012-2016.



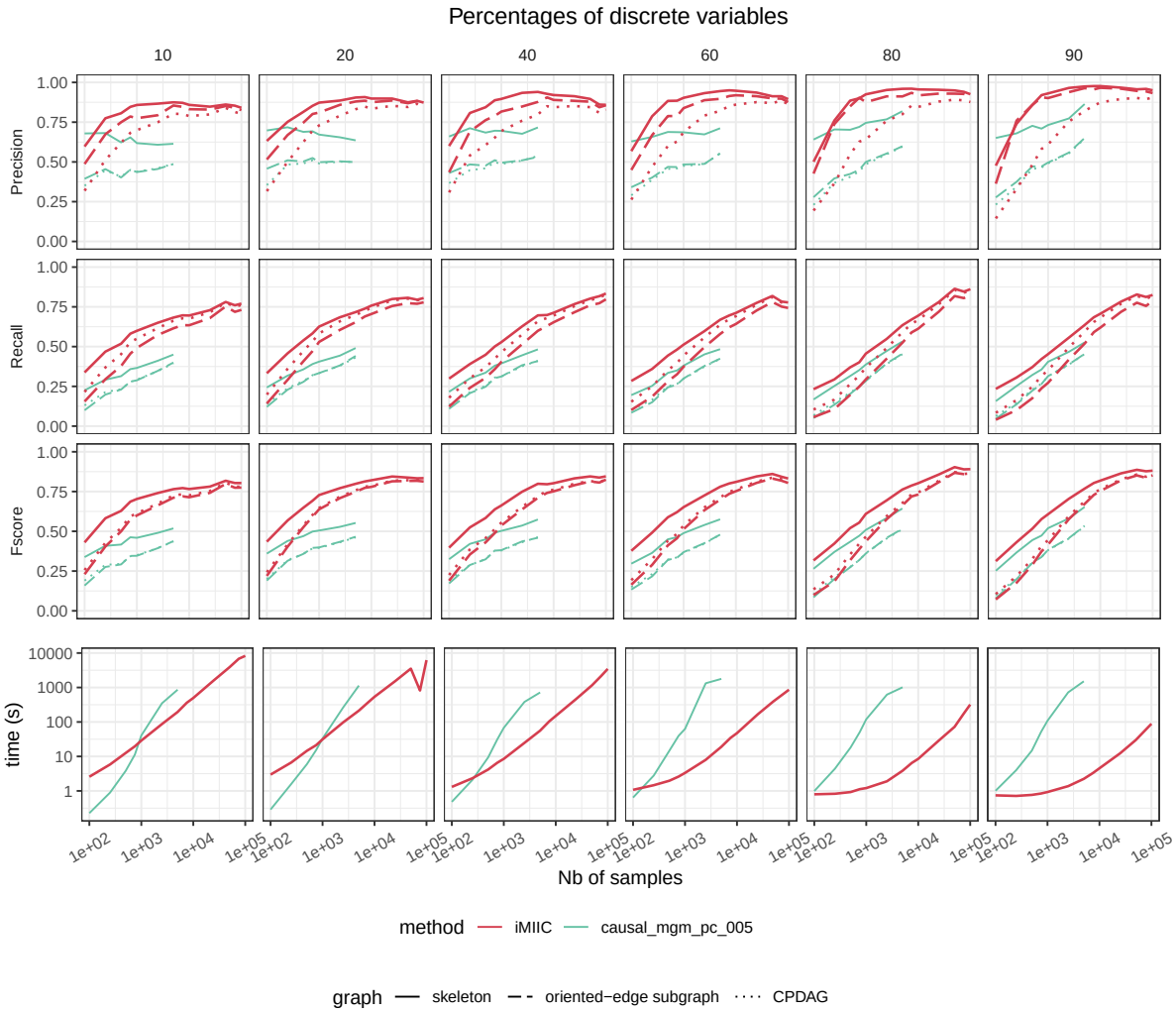

Extended Data Fig. 2: **Example of simulated SEER-like dataset.** Example of marginal distributions of simulated SEER-like datasets (including about 60% of discrete variables here) obtained using mixed-type structural equation models (SEMs)[8], see Data generation and benchmarks in Methods.



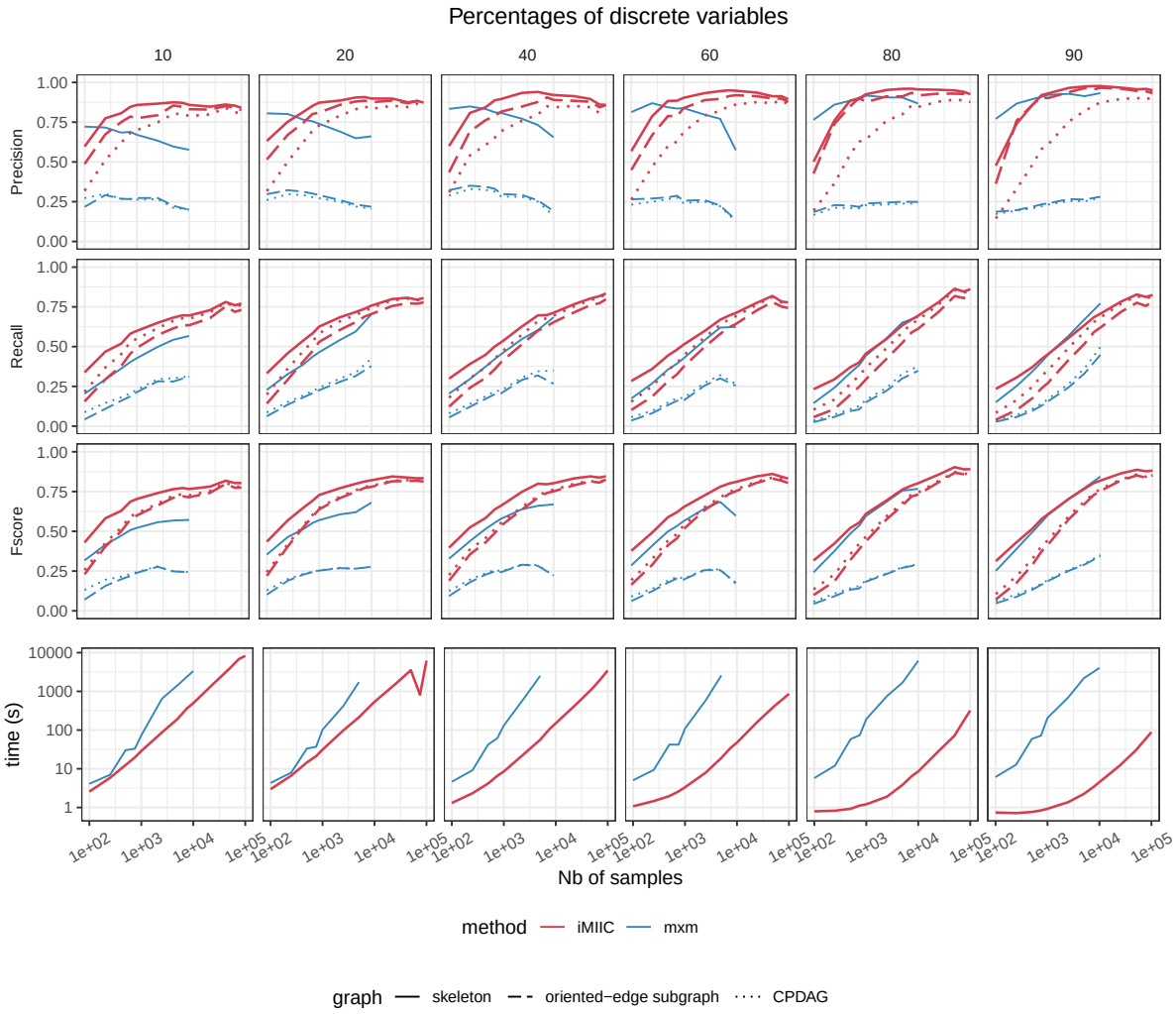

Extended Data Fig. 3: **Original MIIC *versus* PC on SEER-like benchmarks.** See parameter settings in Data generation and benchmarks in Methods. Oriented-edge subgraph scores (dashed lines) are restricted to the subgraphs containing only oriented edges in the theoretical CPDAG *versus* the inferred graph. These oriented-edge scores are designed to specifically assess the method performance on causal discovery, that is, on the oriented edges which can in principle be learnt from observational data *versus* those effectively predicted by the causal structure learning method.



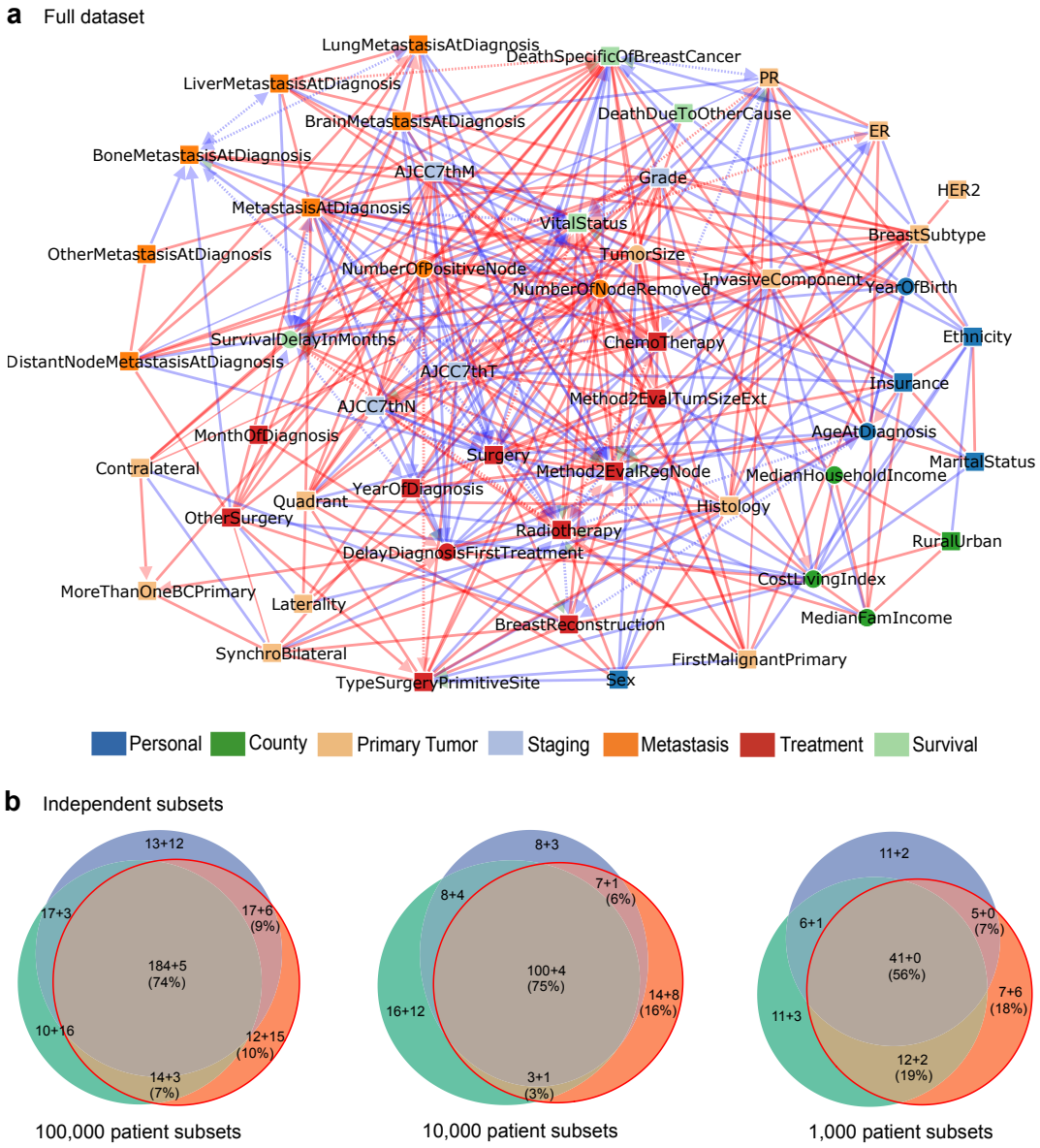

Extended Data Fig. 4: **iMIIC *versus* PC on SEER-like benchmarks.** See parameter settings in Data generation and benchmarks in Methods. Oriented-edge subgraph scores (dashed lines) are restricted to the subgraphs containing only oriented edges in the theoretical CPDAG *versus* the inferred graph. These oriented-edge scores are designed to specifically assess the method performance on causal discovery, that is, on the oriented edges which can in principle be learnt from observational data *versus* those effectively predicted by the causal structure learning method.



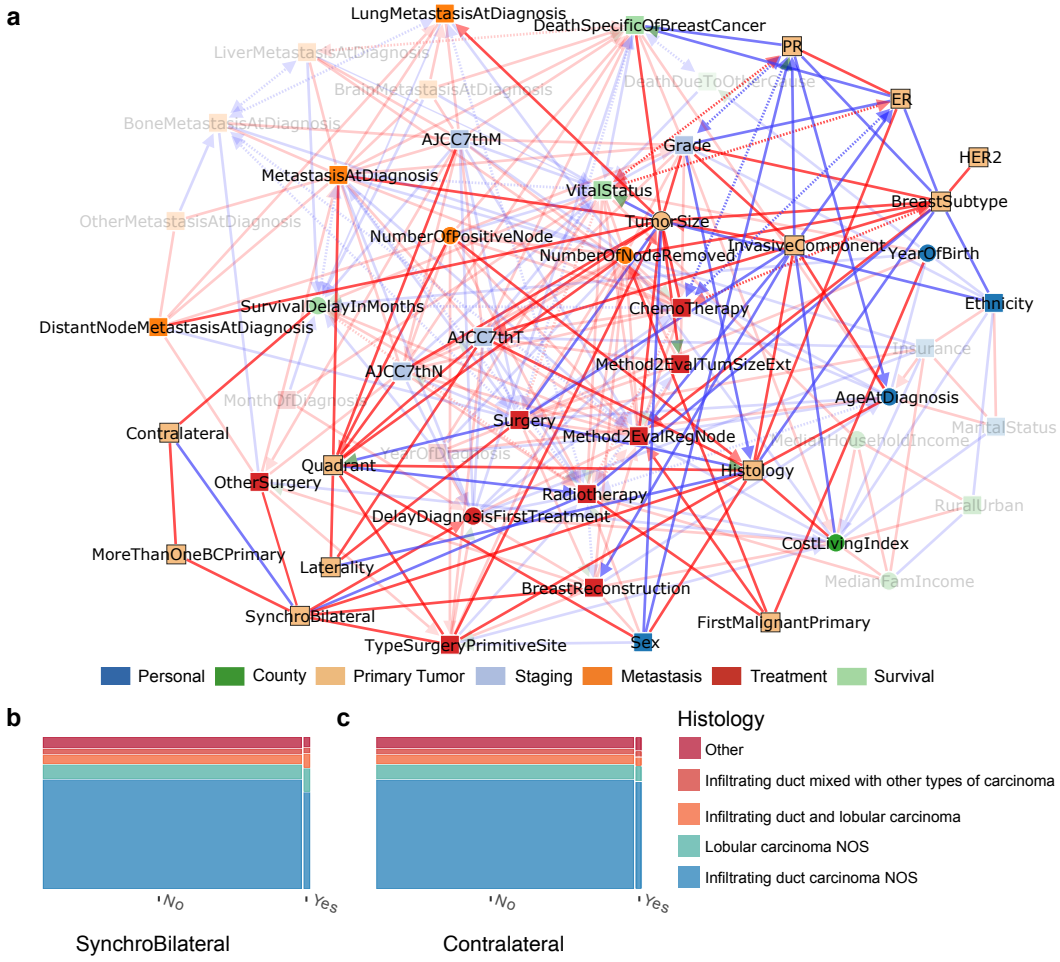

Extended Data Fig. 5: **iMIIC *versus* causalMGM on SEER-like benchmarks.** See parameter settings in Data generation and benchmarks in Methods. Oriented-edge subgraph scores (dashed lines) are restricted to the subgraphs containing only oriented edges in the theoretical CPDAG *versus* the inferred graph. These oriented-edge scores are designed to specifically assess the method performance on causal discovery, that is, on the oriented edges which can in principle be learnt from observational data *versus* those effectively predicted by the causal structure learning method.




Extended Data Fig. 6: **iMIIC *versus* MXM on SEER-like benchmarks.** See parameter settings in Data generation and benchmarks in Methods. Oriented-edge subgraph scores (dashed lines) are restricted to the subgraphs containing only oriented edges in the theoretical CPDAG *versus* the inferred graph. These oriented-edge scores are designed to specifically assess the method performance on causal discovery, that is, on the oriented edges which can in principle be learnt from observational data *versus* those effectively predicted by the causal structure learning method.




Extended Data Fig. 7: **SEER breast cancer orientation consistent networks inferred by iMIIC**. **(a)** The 51 node network inferred by iMIIC from SEER dataset containing 396,179 breast cancer patients diagnosed between 2010 and 2016. This orientation consistent network contains 340 edges and includes 2 contextual variables, Sex and Year of birth. See Supplementary Table 1 for a list and causal nature of each edges predicted by iMIIC. **(b)** Comparisons of networks inferred from three independent subsamplings of the same size of 100,000, 10,000 or 1,000 patient subsets (from left to right). Number of shared edges (regardless of orientations) in the Euler diagrams are given as a sum $a + b$ where $a$ (resp. $b$) corresponds to the number of edges included in (resp. absent from) the full dataset network in (a). Percentages refer to the subset network with the median total number of edges (red circle).




Extended Data Fig. 8: **Primary Tumor subnetwork inferred by iMIIC from SEER breast cancer dataset**. **(a)** Subnetwork highlighting direct relations with primary tumor variables (Contralateral, MoreThanOneBCPrimary, SynchroBilateral, Laterality, Quadrant, Histology, FirstMalignantPrimary, TumorSize, InvasiveComponent, PR, ER, HER2, and BreastSubtype). **(b)** Joint distribution of Histology and Synchro Bilateral tumor. **(c)** Joint distribution of Histology and Contralateral tumor, see main text.

Chapter 6

# Reliable causal discovery based on mutual information supremum principle for finite dataset

> "Resemblance [...] Contiguity [...] and Causation [...] are the only ties of our thoughts, they are really to us the cement of the universe, and all the operations of the mind must, in great measure, depend on them."
> David Hume (1872–1939)

This chapter includes a contribution of this thesis which is a manuscript published, and presented through a poster, in the WHY 21 workshop that took place during the 2021 Conference on Neural Information Processing Systems (NeurIPS 2021).

# Reliable causal discovery based on mutual information supremum principle for finite datasets

**Vincent Cabeli, Honghao Li, Marcel da Câmara Ribeiro-Dantas, Franck Simon, Hervé Isambert**
Institut Curie, Université PSL, Sorbonne Université,
CNRS UMR168, 75005 Paris, France
first-name.last-name@curie.fr

## Abstract

The recent method, MIIC (Multivariate Information-based Inductive Causation), combining constraint-based and information-theoretic frameworks, has been shown to significantly improve causal discovery from purely observational data. Yet, a substantial loss in precision has remained between skeleton and oriented graph predictions for small datasets. Here, we propose and implement a simple modification, named conservative MIIC, based on a general mutual information supremum principle regularized for finite datasets. In practice, conservative MIIC rectifies the negative values of regularized (conditional) mutual information used by MIIC to identify (conditional) independence between discrete, continuous or mixed-type variables. This modification is shown to greatly enhance the reliability of predicted orientations, for all sample sizes, with only a small sensitivity loss compared to MIIC original orientation rules. Conservative MIIC is especially interesting to improve the reliability of causal discovery for real-life observational data applications.

## 1 Background

Constraint-based structure learning methods can, in principle, discover causal relations in purely observational data (Pearl, 2009; Spirtes, Glymour, and Scheines, 2000). This is theoretically feasible up to some independence equivalence classes, as the orientations of certain edges may only be uncovered through perturbative data and remain undetermined if only observational data is available. Yet, regardless of this theoretical limitation, it has long been recognized (Ramsey, Spirtes, and Zhang, 2006; Colombo and Maathuis, 2014) that orientations predicted by constraint-based methods are often unreliable, which has largely limited, in practice, the application of constraint-based methods to uncover causal relations in real-life observational data.

This causal uncertainty originates from the extensive number of steps and conditions that constraint-based methods, such as the original IC (Pearl and Verma, 1991) and PC (Spirtes and Glymour, 1991) algorithms, have to meet before they can infer edge orientation. Indeed, they must first learn an undirected skeleton, by uncovering (conditional) independences between all pairs of variables, before inferring the orientation of v-structures and finally propagating these orientations to other undirected edges. This long chain of uncertain computational decisions leads to the accumulation of errors which ultimately limit the accuracy of the final orientation and propagation steps of constraint-based methods. As a result, edge orientations significantly reduce the precision (or positive predicted value) of inferred causal graphs compared to their undirected skeleton. In addition, constraint-based methods are known to suffer from much lower sensitivity or recall (*i.e.*, true positive rate) than precision scores, in general (Colombo and Maathuis, 2014; Li et al., 2019). This is related to the fact that separating sets used to remove edges in the (early) steps of constraint-based methods are frequently not consistent with the final skeleton and oriented graphs (Li et al., 2019). They correspond to

WHY-21 @ 35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

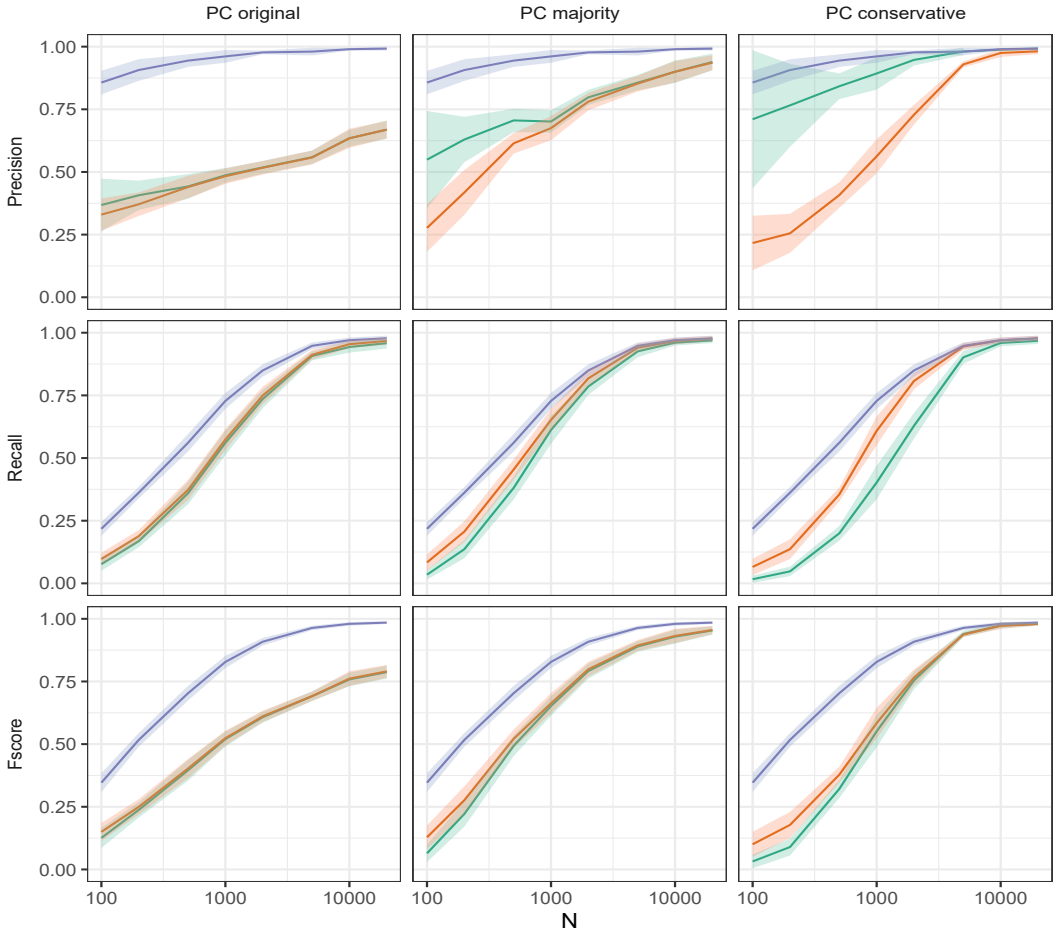



Figure 1: **PC original, majority and conservative orientation rules on discrete datasets**. Benchmark datasets are generated from random 100-node DAGs with average degree 3.8 and maximum degree 4 (See Data generation and benchmarks section for details). PC structure learning performance is measured in terms of Precision, Recall and F-scores ($\pm\sigma$) for skeleton (blue), CPDAG (red) and oriented-edge-only subgraph (green).

spurious conditional independences responsible for the large number of false negative edges and, therefore, low sensitivity of constraint-based methods.

While successive refinements of orientation rules, such as conservative rules (Ramsey, Spirtes, and Zhang, 2006) and majority rules (Colombo and Maathuis, 2014), have helped improve the average precision of orientations, they also lead to large precision variance and further aggravate the poor recall of edge orientations at small sample sizes. This is illustrated here for both discrete (Fig. 1) and continuous (Fig. 2) benchmark datasets generated by random Bayesian networks using the available codes from (Cabeli et al., 2020), see section on Data generation and benchmarks, below.

The recently developed method, MIIC, combining constraint-based and maximum likelihood frameworks, has been shown to significantly improve the situation by greatly reducing the imbalance between precision and recall, for all sample sizes (Verny et al., 2017; Cabeli et al., 2020). Compared to traditional constraint-based methods, MIIC also significantly reduces the precision gap between skeleton and oriented graphs for large enough datasets, as discussed below. However, a substantial loss in precision remains between skeleton and oriented graphs for smaller datasets.

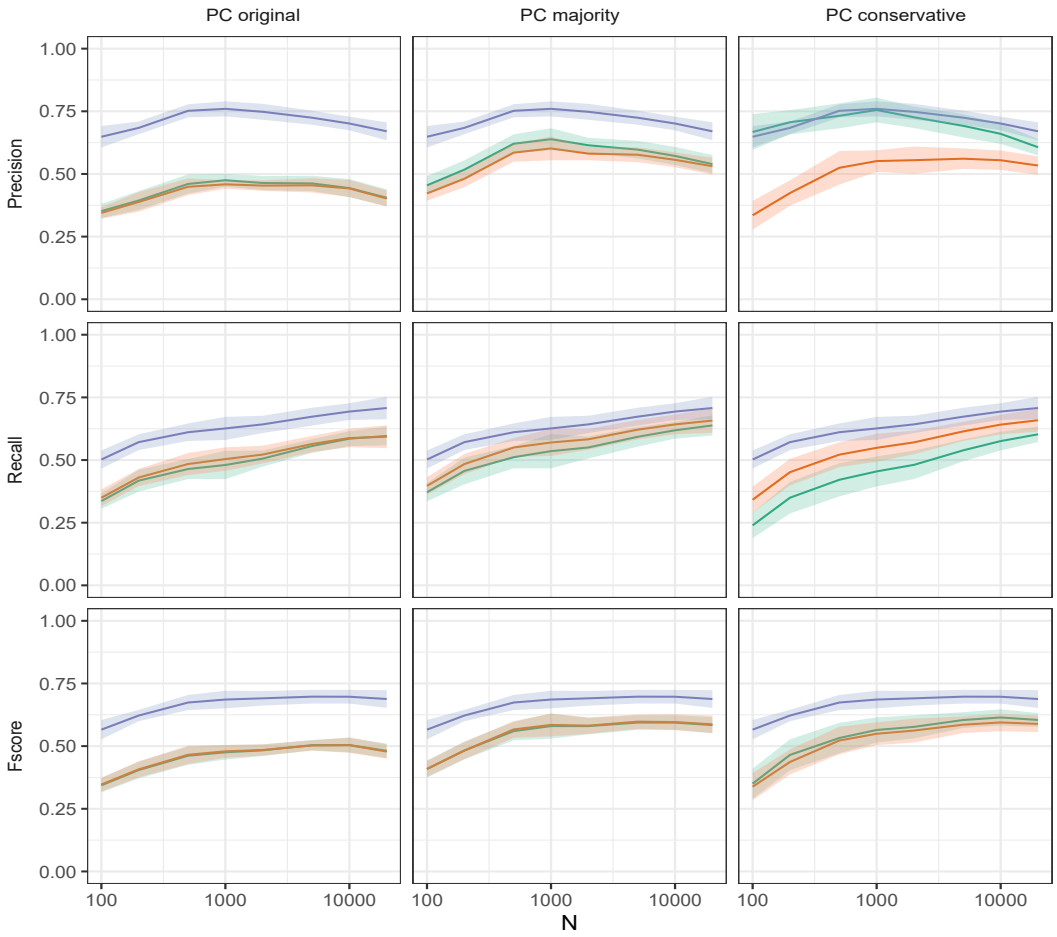



Figure 2: **PC original, majority and conservative orientation rules on continuous datasets**. Benchmark datasets are generated from random 100-node DAGs with average degree 3.8 and maximum degree 4 (See Data generation and benchmarks section for details). PC structure learning performance is measured in terms of Precision, Recall and F-scores ($\pm\sigma$) for skeleton (blue), CPDAG (red) and oriented-edge-only subgraph (green).

In this paper, we propose and implement a simple modification of MIIC algorithm, which is found to greatly improve the precision of predicted orientations even for relatively small datasets. It is achieved at the expense of a small loss of orientation recall but significantly enhances the reliability of predicted orientations for all sample sizes. This simple modification, referred to as conservative MIIC, is especially interesting, in practice, to improve the reliability of causal discovery for real-life observational data applications.

## 2 Results

### 2.1 MIIC outline

MIIC (Multivariate Information-based Inductive Causation) is a novel structure learning method (Verny et al., 2017; Cabeli et al., 2020) and online server (Sella et al., 2018), combining constraint-based and information-theoretic frameworks. Starting from a fully connected graph, MIIC iteratively removes dispensable edges, by uncovering significant information contributions from indirect paths based on the "3off2" scheme (Affeldt and Isambert, 2015; Affeldt, Verny, and Isambert,

2016). This amounts to progressively uncover the best supported conditional independencies, *i.e.* $I(X;Y|\{A_i\}_n) \simeq 0$, by iteratively "taking off" the most significant indirect contributions of *positive* conditional 3-point information, $I(X;Y;A_k|\{A_i\}_{k-1}) > 0$, from every 2-point (mutual) information, $I(X;Y)$, as,

$$I(X;Y|\{A_i\}_n) = I(X;Y) - I(X;Y;A_1) - I(X;Y;A_2|A_1) - \cdots - I(X;Y;A_n|\{A_i\}_{n-1}) \quad (1)$$

In practice, (conditional) independence is established by comparing mutual information (MI) or conditional mutual information (CMI) to a universal Normalized Maximum Likelihood (NML) complexity term, $k_N^{\mathrm{NML}}(X;Y|\{A_i\})/N$, computed over all datasets of the same size $N$ and marginal distributions $p(X,\{A_i\})$ and $p(Y,\{A_i\})$ (Affeldt and Isambert, 2015). This can be seen as a NML-regularization of MI and CMI for datasets of finite sample size $N$ as,

$$I'_N(X;Y|\{A_i\}) = I_N(X;Y|\{A_i\}) - \frac{1}{N}\, k_N^{\mathrm{NML}}(X;Y|\{A_i\}) \quad (2)$$

where $k_N^{\mathrm{NML}}(X;Y|\{A_i\})$ is computed iteratively in linear time (Kontkanen and Myllymäki, 2007; Roos et al., 2008) for increasing numbers of $X$ and $Y$ partitions, $r_x$ and $r_y$, starting with $k_N^{\mathrm{NML}}(X;Y|\{A_i\}) = 0$ for $r_x = r_y = 1$ (Affeldt and Isambert, 2015; Cabeli et al., 2020).

Hence, (conditional) independence is established for $I'_N(X;Y|\{A_i\}) \leqslant 0$, whenever sufficient and significant indirect positive contributions could be iteratively collected in Eq. 1 to warrant the removal of the $XY$ edge.

This leads to an undirected skeleton, which MIIC then (partially) orients based on the sign and amplitude of the NML-regularized conditional 3-point information terms (Affeldt and Isambert, 2015; Verny et al., 2017), corresponding to the difference between NML-regularized (C)MI terms.

$$I'_N(X;Y;Z|\{A_i\}) = I'_N(X;Y|\{A_i\}) - I'_N(X;Y|\{A_i\},Z) \quad (3)$$

In particular, negative NML-regularized conditional 3-point information terms, $I'_N(X;Y;Z|\{A_i\}) < 0$, correspond to the signature of causality in observational data (Affeldt and Isambert, 2015) and lead to the prediction of a v-structure, $X \to Z \leftarrow Y$, if $X - Z - Y$ is an unshielded triple in the skeleton (with $I'_N(X;Y|\{A_i\}) \leqslant 0$). By contrast, a positive NML-regularized conditional 3-point information term, $I'_N(X;Y;Z|\{A_i\}) > 0$, suggests to propagate the orientation of a previously directed edge $X \to Z - Y$ as $X \to Z \to Y$ (with $I'_N(X;Y|\{A_i\},Z) \leqslant 0$), to fulfill the assumptions of the underlying graphical model class.

### 2.2 MIIC performance on discrete data, allowing for negative NML-regularized MI & CMI

MIIC was originally developed for discrete variables only, for which MI and CMI are straightforward to compute. Compared to traditional constraint-based methods on discrete data, MIIC greatly reduces the imbalance between precision and recall, for all sample sizes, Fig. 3. MIIC also significantly reduces the precision gap between skeleton and oriented graphs, for large enough datasets. However, a substantial loss in precision remains between skeleton and oriented graphs, for small datasets, irrespective of the CPDAG or oriented-edge-only subgraph scores used for the comparison, Fig. 3.

These results illustrate the interest in integrating multivariate information criteria into constraint-based methods. However, for small datasets or datasets including variables with many discrete levels, NML complexities can easily out-weight MI and CMI terms for weakly dependent variables. As a result, MIIC tends to infer some v-structure orientations, $X \to Z \leftarrow Y$, for which both NML-regularized (C)MI terms in Eq. 3 are negative, *i.e.* $I'_N(X;Y|\{A_i\}) < I'_N(X;Y|\{A_i\},Z) < 0$, suggesting that $Z$ could in fact be included in a separating set of the $\{X,Y\}$ pair, in contradiction with the inferred v-structure, $X \to Z \leftarrow Y$.

Note that such a v-structure would be excluded from the final graph in the frame of traditional constraint-based methods implementing conservative orientation rules, which check that $Z$ is not included in any separating set of the $\{X,Y\}$ pair (Ramsey, Spirtes, and Zhang, 2006). Similarly, rectifying all negative NML-regularized (C)MI values into null values, as proposed and implemented in the present paper below, leads to a vanishing NML-regularized (conditional) 3-point information term in Eq. 3, *i.e.* $I'_N(X;Y;Z|\{A_i\}) = 0$, which precludes the orientation of the unshielded triple, $X - Z - Y$.

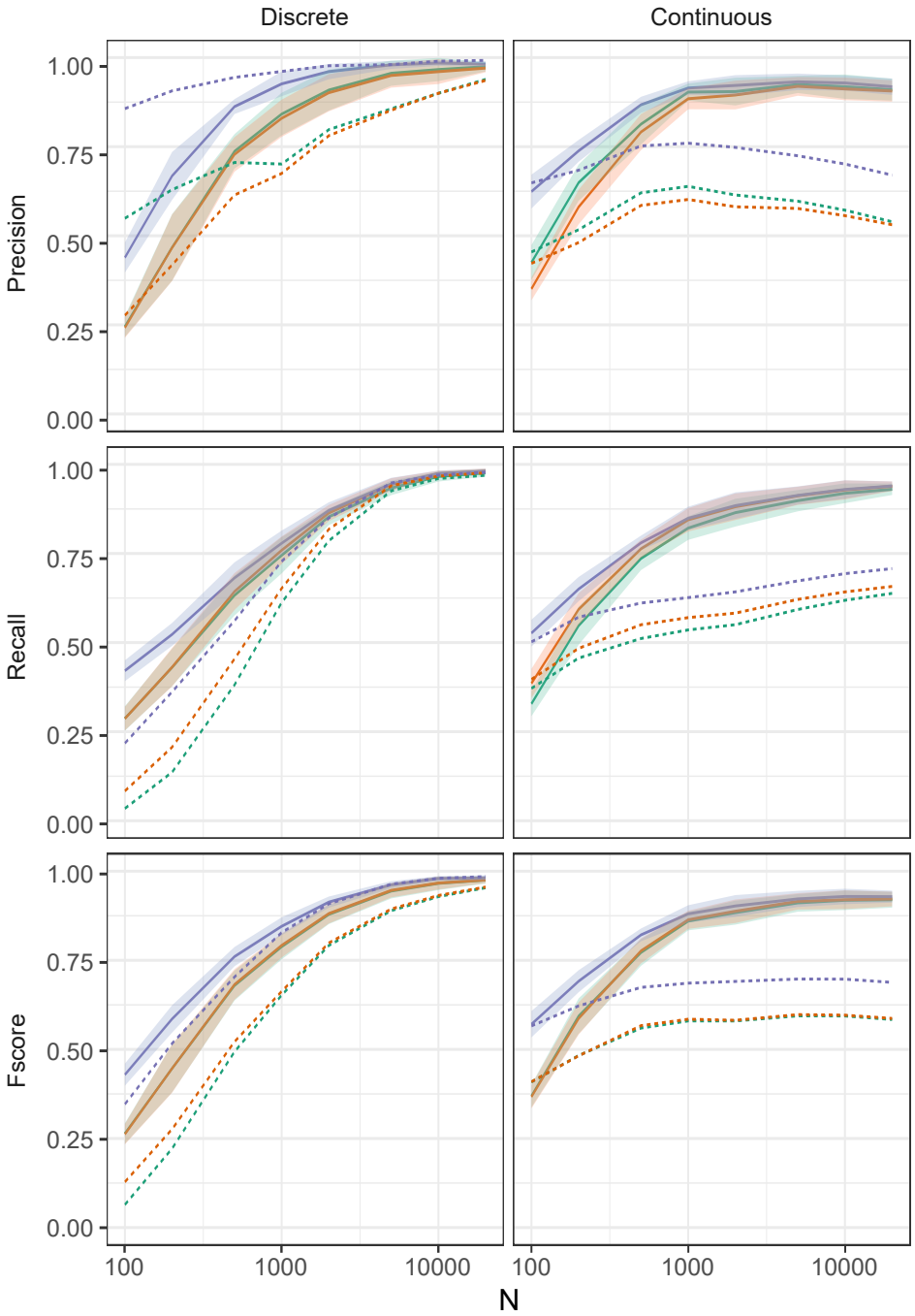



Figure 3: **Original MIIC with orientation rules allowing for negative NML-regularized MI & CMI on discrete data (left) and negative NML-regularized CMI on continuous data (right)**. Benchmark datasets are the same as in Figs. 1 & 2. MIIC structure learning performance is measured in terms of Precision, Recall and F-scores ($\pm\sigma$) for skeleton (blue), CPDAG (red) and oriented-edge-only subgraph (green). PC average scores for majority orientation rules are shown as dashed lines for comparison.

### 2.3 MIIC performance on continuous data, allowing for negative NML-regularized CMI

More recently MIIC was extended to handle continuous as well as mixed-type variables (either combination of discrete and continous variables or variables with both continuous and discrete ranges of values), for which MI & CMI are notoriously more difficult to estimate (Cabeli et al., 2020).

While distance-based k-nearest neighbor (kNN) estimates of MI and CMI are often used for continuous variables (Kraskov, Stögbauer, and Grassberger, 2004; Frenzel and Pompe, 2007), MIIC's MI and CMI estimates are instead computed through an approximate optimum discretization scheme, based on a general MI supremum principle (Cover and Thomas, 2006) regularized for finite datasets and using an efficient $\mathcal{O}(N^2)$ dynamic programming algorithm (Cabeli et al., 2020). This approach finds optimum partitions, $\mathcal{P}$ and $\mathcal{Q}$, specifying the number and positions of cut-points of each continuous variable, $X$ and $Y$, to maximize the NML-regularized MI between them,

$$I'_N(X;Y) = \sup_{\mathcal{P},\mathcal{Q}} I'_N([X]_\mathcal{P};[Y]_\mathcal{Q}) \tag{4}$$

The NML regularization term, introduced in $I'_N([X]_\mathcal{P};[Y]_\mathcal{Q})$, is necessary for finite datasets and amounts to a model complexity cost, which eventually out-weights the information gain in refining bin partitions further, when there is not enough data to support such a refined model (Cabeli et al., 2020).

Such optimization-based estimates of MI are at par with alternative distance-based kNN approaches but have also the unique advantage of providing an effective independence test to identify independent continuous or mixed-type variables (Cabeli et al., 2020). This is achieved when partitioning $X$ and $Y$ into single bins maximizes the NML-regularized MI in Eq. 4, which vanishes exactly, in this case, with dramatic reductions in sampling error and variance (Cabeli et al., 2020). By contrast, kNN-MI estimates still need an actual independence test to decide whether some variables are effectively independent or not, as kNN MI estimates are never exactly null.

MIIC Precision, Recall and F-score on continuous data are comparable to those on discrete data, Fig. 3, and typically much better than the results obtained with traditional constraint-based methods, which, unlike MIIC, need to rely on independence tests, that are notoriously difficult for continuous data.

However, by contrast with discrete data, the remaining loss between skeleton and oriented graph precisions appears to differ between the CPDAG score and the oriented-edge-only subgraph score used for the comparison, Fig. 3. It indicates that the precision of the oriented-edge-only subgraph is slightly though significantly better than for the overall partially oriented graph, with a small concomitant loss of orientation recall, at small sample sizes, Fig. 3. This trend is due to the more stringent condition for v-structure orientation brought by the non-negative NML-regularized MI estimates obtained by MIIC for continuous variables. Yet, the optimum partitioning principle only applies to MI (Cover and Thomas, 2006), not CMI, which need to be estimated through the *difference* between optimum NML-regularized MI terms, as $I'_N(X;Y|U) = I'_N(Y;\{X,U\}) - I'_N(Y;U) = I'_N(X;\{Y,U\}) - I'_N(X;U)$ (Cabeli et al., 2020). As a result, the approximate NML-regularized CMI estimates between conditionally independent variables can sometime be negative and lead to v-structure orientations contradicting conditional independence, as discussed for discrete data above.

### 2.4 Improving MIIC causal discovery by rectifying negative NML-regularized MI & CMI

The general MI supremum principle (Cover and Thomas, 2006), regularized in Eq. 4 for finite datasets, is theoretically valid for any type of variables, not just continuous variables. In particular, it could be applied to small size datasets with discrete or categorical variables with many levels. It would result in the merging of rare levels to better estimate MI and CMI between weakly dependent discrete variables. Ultimately, MI estimates between independent discrete variables should lead to the merging of each variable into a single bin, thereby, resulting in NML-regulatized MI estimates to vanish exactly in this case, as already observed for continuous variables (Cabeli et al., 2020). As a result, optimum NML-regulatized MI should be non-negative as well as, by extension, NML-regulatized CMI, as shown now.

**Theorem 1.** *Optimum NML-regulatized MI and NML-regulatized CMI are non-negative.*

*Proof.* We first address optimum NML-regularized MI, noting that $I'_N(X;Y) \geqslant I'_N([X]_1;[Y]_1) = 0$, where $[X]_1$ and $[Y]_1$ are the $X$ and $Y$ variables partitioned into single bins, which leads to a vanishing

NML-regularized MI, as both MI and NML complexity cost are null, in this case, as $k_N^{\mathrm{NML}}(X;Y) = 0$ for $r_x = r_y = 1$ (Affeldt and Isambert, 2015).

Then, NML-regularized CMI is defined as the *difference* between optimum NML-regularized MI terms as, $I'_N(X;Y|U) = I'_N(Y;\{X,U\}) - I'_N(Y;U) = I'_N(X;\{Y,U\}) - I'_N(X;U)$. However, partitioning $X$ and $Y$ into a single bin leads to $I'_N(Y;\{X,U\}) \geqslant I'_N(Y;\{[X]_1,U\}) = I'_N(Y;U)$ and $I'_N(X;\{Y,U\}) \geqslant I'_N(X;\{[Y]_1,U\}) = I'_N(X;U)$ thus implying $I'_N(X;Y|U) \geqslant 0$ □

Following these considerations on the negativity of NML-regularized (C)MI with MIIC original orientation implementation, we propose a small modification, based on Theorem 1 and referred to as conservative MIIC, by analogy to the conservative orientation rules of traditional constraint-based methods (Ramsey, Spirtes, and Zhang, 2006), as noted above.

**Proposition 2.** *Conservative MIIC rectifies negative values of NML-regularized (C)MI, indicating (conditional) independence, to null values instead.*

The effects on this modification on discrete and continuous benchmark data are show in Fig. 4. While conservative MIIC hardly affects skeleton scores, it clearly has an impact on CPDAG and oriented-edge-only subgraph scores, which exhibit different trends relative to their original MIIC values.

CPDAG Precision, Recall and, hence, F-scores appear to be slightly lower under conservative MIIC (Fig. 4) than with original MIIC (Fig. 3), for discrete data. This illustrates the overall "better" orientation/non-orientation scores of the original MIIC against the theoretical CPDAG objective. Indeed, allowing for negative NML-regularized MI enables to infer weakly supported v-structures at small sample sizes. Besides, no significant difference is observed for CPDAG scores on continuous data, as original MIIC already enforces non-negative NML-regularized MI through optimization for continuous data (Cabeli et al., 2020), suggesting that enforcing also non-negative NML-regularized CMI with conservative MIIC has little impact on the reliability of CPDAG scores for continuous data, at least for the benchmarks tested here.

By contrast, conservative MIIC is found to greatly improve the precision of oriented-edge-only subgraphs, on discrete datasets, even for relatively small sample sizes, Fig. 4. This large increase in orientation precision is achieved at the expense of a relatively small loss of orientation recall. Hence, conservative MIIC significantly enhances the reliability and sensitivity of predicted orientations for all sample sizes, as compared to traditional constraint-based methods with conservative orientation rules, Fig. 4. For instance, conservative MIIC already reaches nearly 90% orientation precision with 25% orientation recall for $N \simeq 250$ (against about 80% orientation precision with only 5% orientation recall for conservative PC). While, by the time conservative PC reaches 90% orientation precision with 25% orientation recall for $N \simeq 700$, conservative MIIC achieves nearly 100% orientation precision with 50% orientation recall, Fig. 4. In addition, while original MIIC achieves a significantly better 65% orientation recall for $N \simeq 700$, Fig. 3, its orientation precision simultaneously drops to about 75%, which clearly impacts its reliability for causal discovery.

On continuous data, conservative MIIC also achieves a large increase in orientation precision, which becomes at par with skeleton precision, even for small datasets, and clearly much better than the corresponding scores obtained with traditional constraint-based methods for large datasets, Fig. 4. For instance, conservative MIIC reaches nearly 75% orientation precision with 50% orientation recall for $N \simeq 200$ (against about 70% orientation precision with 35% orientation recall for conservative PC). While, by the time conservative PC reaches 75% orientation precision with 45% orientation recall for $N \simeq 1,000$, conservative MIIC achieves more than 90% orientation precision with 80% orientation recall, Fig. 4.

## 3 Data generation and benchmarks

Datasets were simulated using structural equations models (SEMs) following the causal order of randomly generated DAGs. Continuous examples were constructed using linear and non-linear functions, and discrete datasets using unique state probabilities for each of the parents' combinations. The DAGs themselves were randomly drawn from the space of all possible 100 node DAGs (Melancon and Philippe, 2004) allowing for a maximum degree of 4 neighbors, resulting in an average degree of 3.8. Further details and dataset examples can be found in Cabeli et al. (2020).

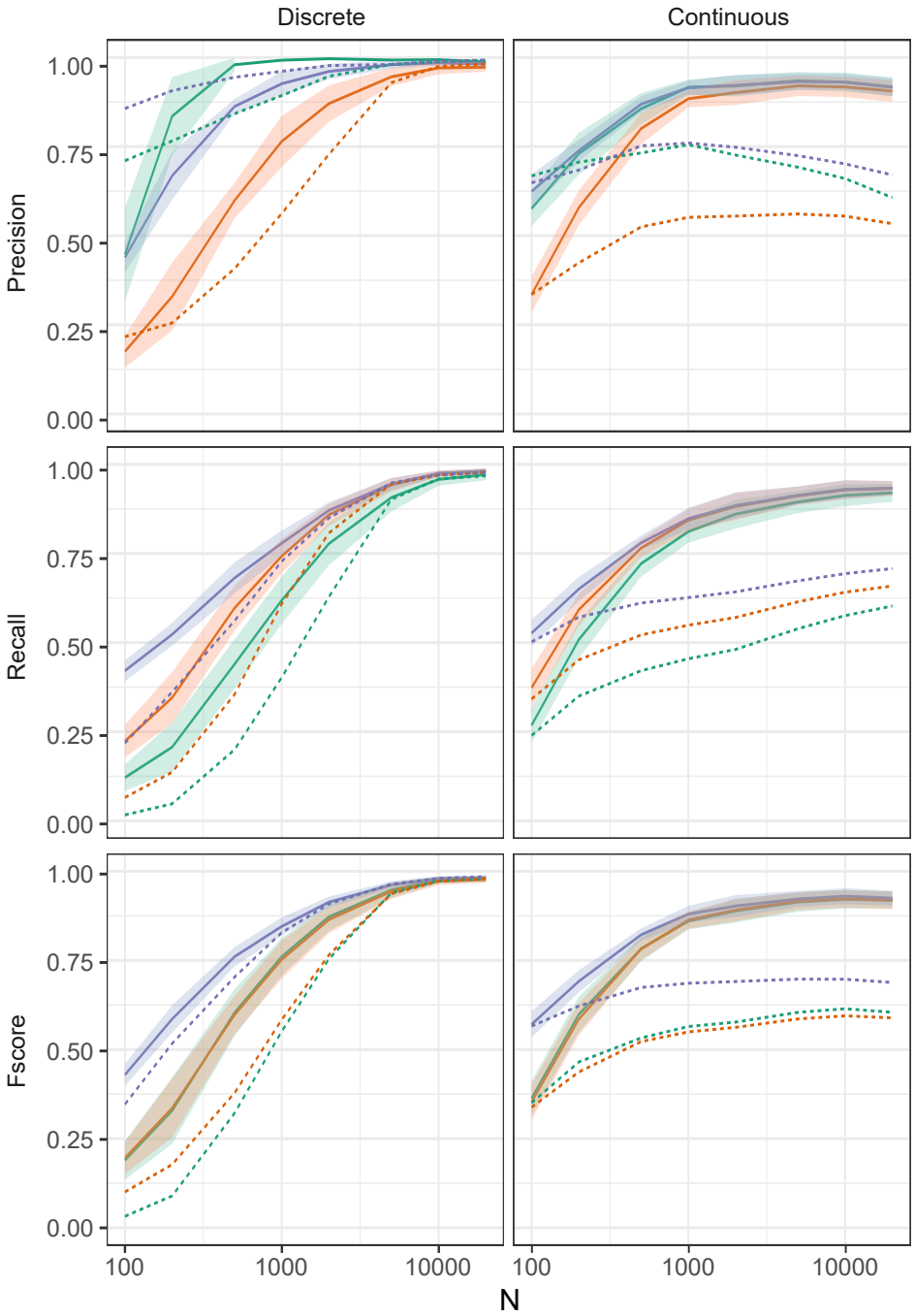



Figure 4: **Conservative MIIC with new orientation rules enforcing non-negative NML-regularized MI & CMI on discrete data (left) as well as continuous data (right)**. Benchmark datasets are the same as in Figs. 1 & 2. Conservative MIIC structure learning performance is measured in terms of Precision, Recall and F-scores ($\pm\sigma$) for skeleton (blue), CPDAG (red) and oriented-edge-only subgraph (green). PC average scores for conservative orientation rules are shown as dashed lines for comparison.

For evaluation purposes, network reconstruction was treated as a binary classification task and classical performance measures, Precision, Recall and F-score, were first used to evaluate skeleton reconstruction, based on the numbers of true *versus* false positive ($TP$ *vs* $FP$) edges and true *versus* false negative ($TN$ *vs* $FN$) edges, irrespective of their orientation.

Then, in order to evaluate edge orientations, we also define two orientation-dependent measures.

The first measure, referred to as the "CPDAG" score, aims to score the overall reconstruction with regards to the equivalence class of the true DAG. Edge types are used to redefine the orientation-dependent counts as, $TP' = TP - TP_{misorient}$ and $FP' = FP + TP_{misorient}$ with $TP_{misorient}$ corresponding to all true positive edges of the skeleton with a different orientation/non-orientation status as in the true CPDAG. The CPDAG precision, recall and F-score were then computed with the orientation-dependent $TP'$ and $FP'$. In particular, the CPDAG score equivalently rates as "false positive" the erroneous orientation of an non-oriented edge in the CPDAG and the erroneous non-orientation of an oriented edge in the CPDAG. However, these errors are not equivalent from a causal discovery perspective.

The second measure, referred to as oriented-edge-only score, uses the same metrics but is restricted to the subgraphs of the CPDAG and the inferred graph containing oriented edges only. It is designed to specifically assess the method performance with regards to causal discovery, that is, on the oriented edges which can in principle be learnt from observational data *versus* those effectively predicted by the causal structure learning method.

MIIC was run with default parameters for all settings on the latest version (available at `https://github.com/miicTeam/miic_R_package`), and PC with the `pcalg` package (Kalisch et al., 2012) using `bnlearn`'s (Scutari, 2010) mutual information test for discrete datasets and rank correlation for continuous ones. For PC, the $\alpha$ threshold for significance testing was tuned for each sample size $N$ and network type to produce the best average between skeleton and "CPDAG" F-scores using a zeroth order optimization implemented in `dlib` (King, 2009).

## 4 Conclusion

Causal uncertainty and limited sensitivity of traditional constraint-based methods have so far hampered their dissemination for a wide range of possible causal discovery applications on real-life observational datasets. Hence, fulfilling the promise of causal discovery methods in the new data analysis area requires to improve their reliability as well as scalability.

We propose and implement, in this paper, a simple modification of the recent causal discovery method, MIIC, which greatly enhances the reliability of predicted orientations, for all sample sizes, with only a small sensitivity loss compared to MIIC original orientation rules. This conservative MIIC approach is especially interesting, in practice, to improve the reliability of cause-effect discovery for real-life observational data applications.

## 5 Acknowledgements

We would like to acknowledge the supports of the following funding agencies: Fondation ARC (VC), French Ministery of Higher Education, Research and Innovation (HL), EU IC-3i cofund PhD programme (MCRD), INSERM ITMO Cancer (FS, HI), CNRS and Institut Curie (HI).

## References

Affeldt, S., and Isambert, H. 2015. Robust reconstruction of causal graphical models based on conditional 2-point and 3-point information. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence (UAI 2015)*, 42–51.

Affeldt, S.; Verny, L.; and Isambert, H. 2016. 3off2: A network reconstruction algorithm based on 2-point and 3-point information statistics. *BMC Bioinformatics* 17(S2):12.

Cabeli, V.; Verny, L.; Sella, N.; Uguzzoni, G.; Verny, M.; and Isambert, H. 2020. Learning clinical networks from medical records based on information estimates in mixed-type data. *PLOS Computational Biology* 16(5):e1007866.

Colombo, D., and Maathuis, M. H. 2014. Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.* 15:3741–3782.

Cover, T. M., and Thomas, J. A. 2006. *Elements of Information Theory*. Wiley, 2nd edition.

Frenzel, S., and Pompe, B. 2007. Partial mutual information for coupling analysis of multivariate time series. *Phys. Rev. Lett.* 99:204101.

Kalisch, M.; Mächler, M.; Colombo, D.; Maathuis, M. H.; and Bühlmann, P. 2012. Causal inference using graphical models with the R package pcalg. *J. Stat. Softw.* 47(11):1–26.

King, D. E. 2009. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research* 10:1755–1758.

Kontkanen, P., and Myllymäki, P. 2007. A linear-time algorithm for computing the multinomial stochastic complexity. *Inf. Process. Lett.* 103(6):227–233.

Kraskov, A.; Stögbauer, H.; and Grassberger, P. 2004. Estimating mutual information. *Phys. Rev. E* 69:066138.

Li, H.; Cabeli, V.; Sella, N.; and Isambert, H. 2019. Constraint-based Causal Structure Learning with Consistent Separating Sets. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, 14257–14266.

Melancon, G., and Philippe, F. 2004. Generating connected acyclic digraphs uniformly at random. *arXiv:cs/0403040*. arXiv: cs/0403040.

Pearl, J., and Verma, T. 1991. A theory of inferred causation. In *In Knowledge Representation and Reasoning: Proc. of the Second Int. Conf.* 441–452.

Pearl, J. 2009. *Causality: models, reasoning and inference*. Cambridge University Press, 2nd edition.

Ramsey, J.; Spirtes, P.; and Zhang, J. 2006. Adjacency-faithfulness and conservative causal inference. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*, UAI, 401–408. Oregon, USA: AUAI Press.

Roos, T.; Silander, T.; Kontkanen, P.; and Myllymäki, P. 2008. Bayesian network structure learning using factorized nml universal models. In *Proc. 2008 Information Theory and Applications Workshop (ITA-2008)*. IEEE Press.

Scutari, M. 2010. Learning bayesian networks with the bnlearn r package.

Sella, N.; Verny, L.; Uguzzoni, G.; Affeldt, S.; and Isambert, H. 2018. Miic online: a web server to reconstruct causal or non-causal networks from non-perturbative data. *Bioinformatics* 34(13):2311–2313.

Spirtes, P., and Glymour, C. 1991. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review* 9:62–72.

Spirtes, P.; Glymour, C.; and Scheines, R. 2000. *Causation, Prediction, and Search*. The MIT Press, Cambridge, Massachusetts, 2nd edition.

Verny, L.; Sella, N.; Affeldt, S.; Singh, P. P.; and Isambert, H. 2017. Learning causal networks with latent variables from multivariate information in genomic data. *PLoS Comput. Biol.* 13(10):e1005662.

## Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] In main text and benchmark figures
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See main text
   (b) Did you include complete proofs of all theoretical results? [Yes] See proof
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The codes for data generation and benchmarks are accessible on github
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Data generation and benchmarks section
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Figures
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Data generation and benchmarks section
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] (Cabeli et al., 2020)
   (b) Did you mention the license of the assets? [Yes] See Data generation and benchmarks section
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See Data generation and benchmarks section
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Chapter 7

# Conclusion

> "A good psychologist has to be able to distinguish strongly between problems of process, which are causal, and problems of structure, which are analytic and descriptive. In particular the statistics adequate for the latter are not sufficient for the former."
> Sir Frederic Charles Bartlett (1886–1969)

This last chapter ends this thesis with overall remarks about the work conducted and presents perspectives and future research.

## 7.1 Interpretable Causal Discovery in Breast Cancer records

To know that correlation does not imply causation, for a long time hindered investigations into the causal nature of natural phenomena. While in some cases questions were simply not asked, in others correlation was escalated to a replacement for causal discussions based some times in fallacies and biased estimates, as presented in Chapter 2. Fortunately, as can be seen in the same chapter, this did not completely prevent attempts to infer causality, and to devise experiments, or take advantage of natural experiments, to understand causes and effects. More recently, it has reached the scientific maturity to even allow us to learn the causal structure based on observational data! iMIIC and iMIIC WebServer, presented in Chapter 3, show great progress in this endeavour providing very informative results about causal relationships, their direction and the degree of belief that we have in them,

such as distinguishing between putative and genuine causal edges.

In Chapter 4, the SEER program was presented in more detail, providing insight of how useful it can be as input for causal discovery methods. The results of all the work developed throughout this thesis can be seen in detail in Chapter 5, where it was shown important results from analyses performed on the final causal graph inferred by iMIIC and displayed by iMIIC WebServer, while in Chapter 6, a conservative version of iMIIC was displayed separately, though it's contained in iMIIC. This thesis points to evidence of how powerful and useful causal discovery methods can be and, with advancements of this hot topic in the scientific literature, how it can contribute to scientific progress in general.

## 7.2 Future Research

iMIIC's benchmarks have proven it to be a top-tier approach for causal discovery for large-scale mixed data. With such confidence in the inferred graph, there is plenty of room to operate causal identification (the first step, verifying if it is possible to quantify the causal effect of an intervention), causal estimation (quantify the causal effect of an intervention) and causal mediation analysis (estimate the partial causal effect due to a mediator). There are open source R and Python packages for many methodologies published in the scientific literature, and it should be viable to make them work together with iMIIC. These new features combined could lift iMIIC to a state of a full and automated causal framework which is unheard of currently.

Appendix A

# Résumé long en français

This appendix includes the *résumé long* (long abstract) that was required in order for the defense to be authorized and was submitted to the doctoral school EDITE de Paris - ED130, L'Ecole Doctorale Informatique, Télécommunications et Electronique.

# Long résumé en français

## Introduction

C'est depuis Platon et Aristote, il y a plus de deux mille ans, qu'on retrouve la trace de discussions sur la causalité. La compréhension de la causalité a évolué dans de nombreuses directions, principalement depuis la révolution scientifique avec les contributions de philosophes tels que David Hume [1]. Depuis lors, avec le développement continu de la statistique, la causalité a suivi principalement deux voies : (a) l'oubli, dû à l'impression, surtout chez les statisticiens, qu'une simple corrélation statistique, telle que le coefficient de corrélation de Pearson (CCP), suffirait à expliquer la relation entre les variables, et (b) une recherche plus poussée sur la façon dont la prise en compte de la causalité pourrait nous aider à mieux comprendre la relation entre les événements. Aujourd'hui, on sait que la connaissance de la causalité peut rendre triviale la compréhension de nombreuses situations qui, par le passé, semblaient paradoxales aux statisticiens, comme le paradoxe de Simpson, le paradoxe de Berkson, le paradoxe de l'insuffisance pondérale à la naissance, le problème de Monty Hall, etc [2].

On sait depuis longtemps que l'apprentissage des relations de cause à effet à partir de données purement observationnelles est, en principe, possible grâce à des travaux fondamentaux sur les méthodes de découverte causale [3, 4]. En substance, la découverte causale infère les relations de cause à effet à partir de schémas de corrélations spécifiques impliquant au moins trois variables, sans contredire la notion populaire selon laquelle la corrélation n'implique pas la causalité, mais en montrant que, dans certaines circonstances particulières, cela pourrait être le cas. Cependant, alors que les données d'observation représentent la grande majorité des données disponibles dans un large éventail de domaines, la découverte des relations de cause à effet reste notoirement difficile en l'absence d'intervention systématique, qui peut être peu pratique, trop coûteuse ou contraire à l'éthique, lorsqu'il s'agit de la santé humaine.

La découverte causale est étroitement liée aux méthodes d'apprentissage de modèles graphiques [3, 4], mais la plupart des méthodes d'apprentissage des structures ne sont pas réellement conçues pour découvrir les relations de cause à effet. En particulier, les approches de maximum de vraisemblance, telles que Search-and-Score [5] ou Lasso graphique [6], sont restreintes à des classes de modèles spécifiques, en supposant soit des graphes entièrement dirigés, soit des graphes entièrement non dirigés, et ne peuvent donc pas apprendre la vraie nature causale ou non causale des données. En revanche, les méthodes de découverte causale basées sur les contraintes supposent des classes de graphes plus larges et peuvent apprendre l'orientation de certaines arêtes uniquement sur la base de données d'observation [3, 4], Fig. 1b. À cette fin, ils apprennent d'abord les contraintes structurelles, sous la forme de relations d'indépendance conditionnelle, qui fournissent des informations indirectes et quelque peu cryptiques sur les relations causales possibles entre les variables observées et non observées, comme indiqué dans l'Encadré 1. Pourtant, bien qu'elles soient théoriquement valables, en faisant l'hypothèse d'une quantité illimité de données [7], les méthodes basées sur les contraintes restent peu fiables et difficiles à interpréter sur les ensembles de données relativement restreints qu'elles traitent en pratique.

Au cours de ce doctorat, une méthode avancée de découverte causale, iMIIC (interpretable MIIC), qui permet d'apprendre des modèles graphiques causaux plus fiables et interprétables, ainsi que de traiter de données beaucoup plus importantes (incluant quelques centaines de milliers d'échantillons) a été développée avec d'autres collaborateurs. La nouvelle méthode iMIIC élargit et améliore considérablement l'interprétabilité et l'évolutivité de la récente méthode d'apprentissage des structures MIIC (Multivariate Information-based Inductive Causation), en combinant une structure basée sur les contraintes et la théorie de l'information [8–10]. En résumé, iMIIC apporte un certain nombre d'avancées, qui améliorent considérablement ses performances en matière de découverte causale sur des jeux de données synthétiques et réels de toutes tailles. En

particulier, iMIIC (i) améliore quantitativement la confiance dans l'orientation des arêtes, (ii) distingue les relations causales "authentiques" des relations causales "putatives" (Encadré 1), (iii) distingue les variables contextuelles des variables stochastiques, (iv) renforce la cohérence des chemins indirects et quantifie leur contribution à l'information et, enfin, (v) permet l'extensibilité à de très grands ensembles de données. Ces capacités accrues, qui reposent sur des avancées conceptuelles et un remaniement algorithmique important, sont appliquées pour reconstruire un réseau causal interprétable à partir de l'analyse de 396 179 dossiers médicaux de patientes atteintes d'un cancer du sein provenant du programme Surveillance, Epidemiology, and End Results (SEER) [11].

Les dossiers médicaux nationaux contiennent des quantités massives de données réelles sur la santé humaine, y compris certaines informations personnelles, familiales et socio-économiques, qui affectent fréquemment non seulement l'état de santé, mais aussi le moment du diagnostic, les traitements médicaux et, la survie des patients. En outre, ces déterminants non médicaux de la santé humaine sont généralement contrôlés dans le cadre d'essais cliniques, qui sélectionnent des groupes spécifiques de patients selon des critères d'inscription restrictifs. De ce fait, la richesse des informations contenues dans les dossiers médicaux réels reste largement sous-exploitée en raison de l'absence de méthodes et d'outils non supervisés permettant de les analyser sans hypothèses préconçues. Cela souligne la nécessité de développer de nouvelles stratégies d'apprentissage automatique pour analyser les données de santé, afin de découvrir des associations insoupçonnées et d'éventuelles relations de cause à effet entre toutes les informations disponibles enregistrées dans les antécédents médicaux des patients.

## Aperçu et limites des méthodes de découverte causale

Les méthodes de découverte causale basées sur les contraintes procèdent par étapes successives, comme le montre la Fig. 1b. La première étape consiste à supprimer, de manière itérative, toutes les arêtes superflues d'un réseau initial entièrement connecté, lorsque deux variables sont indépendantes ou conditionnellement indépendantes compte tenu d'un ensemble dit de séparation des variables de conditionnement. La deuxième étape consiste alors à orienter certaines des arêtes de ce graphe non orienté (nommé squelette) pour former ce que l'on appelle des "structures en V", $X \rightarrow Z \leftarrow Y$, qui sont la signature de la causalité dans les données d'observation, Encadré 1. Enfin, la troisième étape vise à propager les orientations des structures en V aux arêtes en aval, Fig. 1b. Cependant, les méthodes traditionnelles basées sur les contraintes manquent de robustesse sur des ensembles de données finis, car leur longue série de décisions incertaines conduit à une accumulation d'erreurs qui limitent la fiabilité des réseaux finaux. En particulier, les indépendances conditionnelles fallacieuses, provenant de combinaisons coïncidentes de variables de conditionnement, entraînent de nombreuses arêtes faussement négatives, qui, en fin de compte, limitent la précision des orientations déduites. La récente méthode d'apprentissage automatique MIIC [8, 10], apprend des modèles graphiques causaux plus robustes en commençant par collecter itérativement les contributeurs d'informations significatives avant d'évaluer les indépendances conditionnelles. En pratique, la stratégie de MIIC limite les indépendances conditionnelles parasites et améliore considérablement la sensibilité ou le rappel (*i.e.*, la fraction de arêtes correctement récupérées) par rapport aux méthodes traditionnelles basées sur les contraintes. En outre, MIIC peut traiter des données hétérogènes (*i.e.* combinant des variables continues et catégorielles) et des données manquantes [10], ainsi que des variables latentes non observées [8], qui sont omniprésentes dans de nombreuses applications de la vie réelle.

Cependant, la méthode MIIC originale présente encore un certain nombre de limitations, telles qu'une fiabilité moindre dans la prédiction de l'orientation des arêtes par rapport à la présence des arêtes, que la nouvelle méthode iMIIC vise à surmonter, comme indiqué ci-dessous. Dans la pratique, il est démontré que la méthode iMIIC améliore considérablement la fiabilité, l'interprétabilité et l'évolutivité de la découverte causale à partir de données synthétiques à grande échelle, ainsi que d'ensembles de données d'observation réelles.

## Nouvelles caractéristiques avancées de la méthode iMIIC

**iMIIC améliore la fiabilité des orientations déduites.** Alors que MIIC original surpasse de manière significative les méthodes traditionnelles basées sur les contraintes pour déduire des orientations fiables, une perte substantielle de précision subsiste généralement entre les prédictions du squelette MIIC et du graphe orienté. Cela est dû à des erreurs d'orientation provenant principalement de structures en V incohérentes,

$X \to Z \leftarrow Y$, dont le nœud central $Z$ pourrait également être inclus dans l'ensemble de séparation de la paire non connectée $\{X, Y\}$, en contradiction avec la rencontre tête-à-tête de la structure en V. Les structures en V incohérentes sont particulièrement courantes pour les ensembles de données comprenant des variables discrètes avec de (trop) nombreux niveaux. Pour éviter de telles orientations incohérentes, iMIIC met en œuvre des règles d'orientation plus conservatives, basées sur un principe général de supremum d'information mutuelle [16, 17] régularisé pour les ensembles de données finis. En pratique, elle améliore considérablement la fiabilité des orientations prédites avec seulement une petite perte de sensibilité par rapport aux règles d'orientation originales de MIIC, Fig. 1c.

**iMIIC distingue les relations causales "authentiques" des relations "putatives".** Les méthodes traditionnelles basées sur les contraintes et la méthode originale de MIIC ne permettent de découvrir que des relations causales "putatives", car les orientations des structures en V sont en fait compatibles à la fois avec de véritables relations de cause à effet et des effets de causes communes non observées, comme le montre un exemple intuitif dans l'Encadré 1. En revanche, iMIIC distingue les arêtes causales "authentiques" des arêtes causales "putatives" en excluant l'effet d'une cause commune non observée (ou d'un facteur de confusion non mesuré) pour chaque arête causale authentique prédite. Pour ce faire, on évalue les probabilités distinctes de la tête et de la queue de la flèche pour toutes les arêtes orientées. Les véritables arêtes causales (représentées par une tête de flèche verte) sont alors prédites si les probabilités de la tête et de la queue de la flèche sont statistiquement significatives, tandis que les arêtes causales restent "putatives" si leur probabilité de queue n'est pas statistiquement significative ou ne peut être déterminée à partir de données purement observationnelles. De même, les arêtes bidirectionnelles, interprétées comme l'effet de causes communes non observées, correspondent à deux probabilités de tête significatives, tandis que tous les autres cas sont représentés graphiquement comme des arêtes non directionnelles.

**iMIIC distingue les variables contextuelles des variables stochastiques.** Le cadre probabiliste distinct des orientations de la tête et de la queue de la flèche mis en œuvre dans iMIIC permet également d'inclure des connaissances préalables sur certaines orientations de tête ou de queue. Par exemple, l'inclusion de quelques variables contextuelles dans les modèles graphiques peut aider à spécifier un paramètre de contrôle ou des conditions expérimentales ou à caractériser le profil personnel des patients. (*e.g.* sexe, année de naissance), en fonction de la nature de l'ensemble de données. Contrairement à la plupart des autres variables de l'ensemble de données, ces variables contextuelles ne varient pas de façon stochastique et devraient, par hypothèse, n'avoir aucune flèche pointant en leur direction, *i.e.*, $p_{\text{queue}} = 1$. Cela exprime notre connaissance préalable que les variables contextuelles ne peuvent pas être la conséquence d'autres variables observées ou non observées dans l'ensemble de données.

**iMIIC renforce la cohérence des chemins indirects et quantifie leur contribution à l'information.** Le raisonnement qui sous-tend la suppression des arêtes inutiles dans la première étape des méthodes de découverte causale basées sur les contraintes est que toutes les associations statistiques entre les variables déconnectées doivent pouvoir être interprétées graphiquement en termes de chemins indirects dans le réseau final. Cependant, ce n'est pas toujours le cas dans la pratique [18]. En particulier, il n'y a aucune garantie que les ensembles de séparation identifiés au cours de cette suppression itérative d'arêtes restent cohérents en termes de chemins indirects dans le réseau final. À cette fin, iMIIC adapte un nouveau schéma algorithmique [18] pour s'assurer que tous les ensembles de séparation identifiés pour supprimer les arêtes inutiles sont cohérents avec le graphe inféré final. On y parvient en répétant le schéma d'apprentissage de la structure basé sur les contraintes, de manière itérative, tout en ne sélectionnant que les ensembles de séparation qui sont cohérents avec le squelette ou le graphe partiellement orienté obtenu à l'itération précédente, comme indiqué dans la Fig. 1b. Cette cohérence des chemins indirects améliore l'interprétabilité des réseaux inférés par iMIIC en termes d'effets indirects, qui sont également quantifiés par les contributions d'information indirectes.

**iMIIC surpasse les méthodes existantes sur des ensembles de données de référence synthétiques.** Les performances de iMIIC ont été comparées à celles de MIIC original ainsi qu'à d'autres méthodes de pointe basées sur les contraintes sur des ensembles de données de référence avec différentes proportions de variables discrètes. Fig. 1c démontre que iMIIC améliore significativement la précision des orientations au prix d'une perte relativement faible de la sensibilité à l'orientation et du F-score pour des ensembles de données de référence de type SEER avec de grandes proportions de variables discrètes. Par exemple,

pour $N = 500$, la précision de l'orientation (resp. F-score) dépasse déjà 85 % (resp. 32%) avec iMIIC *contre* 73% (resp. 39 %) avec MIIC original, pour des ensembles de données de référence de type SEER avec 80 % variables discrètes, et même 93 % (resp. 25 %) *contre* 64 % (resp. 35 %) pour les ensembles de données entièrement discrets, Fig. 1c. En outre, iMIIC surpasse largement la fiabilité et la sensibilité des orientations déduites par rapport aux autres méthodes de pointe basées sur les contraintes, Fig. 1d. En particulier, les F-score d'orientation de iMIIC sont environ deux fois plus élevés que celles de l'algorithme PC [19, 20], pour toutes les tailles d'échantillon et les proportions de variables discrètes dans ces ensembles de données de type SEER. Par exemple, pour les benchmarks avec 80 % de variables discrètes comme dans l'ensemble de données SEER actuel, iMIIC atteint déjà 88 % (resp. 44 %) en précision (resp. F-score) pour $N = 10^3$, contre environ 60 % (18 %) pour conservative PCx [20, 21], 50 % (36 %) pour causalMGM[22] et 24 % (18 %) pour MXM[23]. Pour $N = 10^4$, iMIIC atteint 92% (73 %) en précision (F-score), contre environ 75% (40 %) pour conservative PC, 62 % (55 %) pour causalMGM et 30 % (30 %) pour MXM. Enfin, iMIIC atteint plus de 90 % tant pour la précision de l'orientation que pour le F-score pour $N = 10^5$, ce qui est au-delà de la taille d'échantillon atteignable par d'autres méthodes.

## Application aux données de dossiers médicaux à l'échelle nationale

**Données SEER sur le cancer du sein.** iMIIC a été appliqué à un vaste ensemble de données sur le cancer du sein [11] du programme Surveillance, Epidemiology, and End Results (SEER) de l'Institut National du Cancer, qui collecte des données sur les diagnostics de cancer, les traitements et la survie pour ∼35 % de la population américaine, Fig. 1a. Le cancer du sein [24] est le cancer invasif le plus fréquent chez la femme et n'est curable que dans 70 à 80 % des cas, avec des patients ayant grandes disparités en termes de sous-types de tumeurs et de stades au moment du diagnostic, de traitement initial et de traitements ultérieurs, ainsi que d'âge, d'origine ethnique, de prédisposition génétique, de mode de vie ou de situation socio-économique. De nombreuses études d'association rétrospectives [25–27] et quelques recherches sur l'inférence causale [28–31] ont été rapportées sur la base de données SEER relative au cancer, ce qui en fait une ressource de référence unique pour évaluer les performances réelles des méthodes de découverte causale sur des données de santé réelles.

**Analyse robuste de découverte causale iMIIC sur ∼400 000 patients atteints de cancer du sein.** L'analyse de découverte causale par iMIIC sur les données du cancer du sein SEER est présentée ici pour la période 2010-2016. Il y a 407 791 dossiers médicaux mais seulement 396 179 patients distincts en raison de tumeurs primaires du sein multiples pour certains patients. Cinquante et une variables cliniques, socio-économiques et de survie ont été sélectionnées pour leur pertinence par rapport au cancer du sein et pour le peu de redondance ou d'information manquante.

Le réseau du cancer du sein qui en résulte, Fig. 2a, fournit un modèle graphique interprétable comprenant 280 arêtes, pour lequel la plupart des relations de cause à effet sont connues ou peuvent être exclues sur la base des connaissances communes ou spécialisées ainsi que de la pratique clinique.

Cette évaluation indique qu'environ 90 % des effets causaux authentiques ou putatifs déduits par iMIIC sont corrects, sur la base des connaissances cliniques et épidémiologiques actuelles, tandis que 8 % supplémentaires des relations de cause à effet semblent plausibles. En outre, aucun des liens causaux authentiques prédits ne relie des paires de variables non spécifiques au cancer, telles que des informations personnelles ou socio-économiques, qui sont susceptibles d'un éventuel biais de sélection [13–15] quant au diagnostic du cancer du sein (Encadré 1). En outre, les facteurs de confusion non mesurés (latents) peuvent être éliminés pour les véritables arêtes causales (Encadré 1), tandis que les contributions des facteurs de confusion mesurés sont estimées comme des contributions indirectes. Pourtant, d'autres sources de biais dans la collecte et l'analyse des données ont été signalées dans la base de données SEER [32, 33] (comme indiqué dans la section suivante). Ce réseau obtenu avec ∼400 000 patients est également robuste au sous-échantillonnage car il comprend 90 % des arêtes de trois réseaux appris à partir de trois sous-ensembles indépendants de 100 000 patients, Fig. 2b. En outre, 88 % des probabilités d'orientation des arêtes sont compatibles entre les trois réseaux des sous-ensembles de 100 000 patients et 92 % de celles-ci sont également compatibles avec les probabilités d'orientation des arêtes du réseau complet.

## Interprétation causale du réseau iMIIC du cancer du sein

Dans cette section, l'interprétation clinique et socio-économique du réseau SEER de cancer du sein déduit par iMIIC est abordée, Fig. 2a. L'accent est mis, en particulier, sur les relations de causalité attendues ou plus

surprenantes mises en évidence par iMIIC, et des interprétations des prédictions cause-effet contre-intuitives seront proposées en termes de parcours de soins, de décisions thérapeutiques, de préférences des patients ou de déterminants socio-économiques des soins de santé. Ces résultats sont présentés du point de vue de différentes classes de variables et de sous-réseaux associés, en commençant par le sous-réseau de survie, puis le sous-réseau de la tumeur primaire, le sous-réseau de la chirurgie et du traitement ultérieur, et enfin le sous-réseau socio-économique.

**Sous-réseau de survie.** Le réseau complet, Fig. 2a, contient quatre nœuds associés au statut de survie des patients à la fin de l'année 2016 et définissant un sous-réseau de survie. Il inclut toutes les variables directement liées au statut de survie des patients, Fig. 3a. Au-delà du statut vital de chaque patient (mort ou vivant), deux nœuds supplémentaires précisent la cause du décès, soit par cancer du sein, soit par toute autre cause, et une troisième variable continue correspond au délai de survie ou de suivi en mois, soumis à la période de censure 2010-2016 de l'étude. Fig. 3a montre que les facteurs connus comme responsables de la mortalité due au cancer du sein sont correctement retrouvés par iMIIC, comme les métastases au moment du diagnostic (taux de mortalité global de 49,2 %), avec les métastases les plus éloignées au moment du diagnostic (cerveau et foie) conservant également des liens directs à la fois avec le décès spécifique au cancer du sein et le statut vital, ce qui explique leur taux de surmortalité, *i.e.* métastases au cerveau (70,5 %) et métastases hépatiques (59,5 %). De même, le nombre de ganglions lymphatiques positifs et les variables de stadification (AJCC7e T, N et M) sont tous correctement reliés à la fois au décès spécifique au cancer du sein et au statut vital, et pas à une autre cause de décès. En revanche, iMIIC déduit des relations causales entre l'année de naissance et le décès dû à une autre cause, ainsi qu'entre l'année de naissance et le statut vital, comme attendu. Il est également possible de constater que le décès des patients, quelle qu'en soit la cause, entraîne à juste titre une réduction de leur délai de survie. Pourtant, la Fig. 3a contient également des résultats moins intuitifs. En particulier, le statut vital semble, de manière robuste, "causalement" conditionner la délivrance d'une radiothérapie, à la fois dans l'ensemble des données et dans les trois sous-ensembles de 100 000 patients, avec 51 % des patients vivants ayant subi une radiothérapie contre seulement 27 % des patients décédés, Fig. 3b. Cela suggère qu'un décès précoce dans les premiers mois après le diagnostic peut empêcher l'accès à la radiothérapie pour certains patients qui auraient pu bénéficier de ce traitement s'ils avaient vécu plus longtemps. Cet effet causal à court terme entre le statut vital et la radiothérapie est cohérent avec le déclin rapide de la distribution du délai de survie pour les 3-6 premiers mois en l'absence de radiothérapie, Fig. 3c, ce qui correspond à la fourchette typique de délais pour la radiothérapie après le diagnostic, selon qu'elle est effectuée secondaire à la chirurgie ou après chirurgie et chimio-thérapie adjuvante [34]. Au total, cet effet causal à court terme du statut vital sur la radiothérapie l'emporte sur l'effet bénéfique, inversement proportionnel, de la radiothérapie sur la survie à long terme des patients. Cela suggère un fort "biais de temps immortel" [32] dans le bénéfice apparent de la radiothérapie, Fig. 3d, qui devrait être corrigé avec la "méthode landark" [32, 35] en excluant les patients décédant au cours d'une période déterminée après l'opération, ou en reproduisant un essai cible à partir de données d'observation [36]. En revanche, la chirurgie - qui est généralement pratiquée dans les 5 à 8 semaines suivant le diagnostic - s'avère être la principale cause du délai de survie prolongé des patients, comme nous le verrons plus loin, Fig. 3e et Fig. 4a.

Enfin, il est à noter qu'un certain nombre de variables qui ont été rapportées comme étant associées aux variables de survie sont en fait indirectement plutôt que directement liées à celles-ci. C'est notamment le cas de l'assurance [37, 38] et l'état matrimonial [39, 40]. L'effet indirect de l'assurance (avec les catégories non assuré / Medicaid / non-Medicaid) sur le décès dû au cancer du sein s'explique indirectement par la chirurgie (50 %), la chimiothérapie (14 %), l'état civil (20 %), la radiothérapie (9 %) et la reconstruction mammaire (7 %). Similairement l'effet indirect de l'état matrimonial (célibataire / marié / séparé / divorcé / veuf) sur ce décès dû au cancer du sein s'explique indirectement par la chirurgie (58 %), l'année de naissance (40 %) et l'origine ethnique (2 %).

**Sous-réseau de la tumeur primaire.** Outre les métastases au moment du diagnostic, le statut des récepteurs hormonaux (ER/PR) et la taille de la tumeur primaire ont également une incidence directe sur le pronostic vital des patientes. En particulier, iMIIC déduit que le statut ER réduit le risque de décès dû au cancer du sein de 17,7 % (ER-) à 5,4 % (ER+), avec une contribution indirecte importante (82 %) du statut PR. Ce résultat est cohérent avec le contrôle transcriptionnel de l'ER de PR [41] et un taux de mortalité significativement plus élevé chez les patientes ER+/PR- (11,8 %) que chez les patientes ER+/PR+ (4,4 %).

De même, iMIIC déduit un certain nombre d'associations directes entre l'histologie des tumeurs primaires et d'autres variables, telles que l'âge au moment du diagnostic (en accord avec les premiers rapports) [42]) et avec des tumeurs primaires bilatérales synchrones (détectées dans les 6 mois suivant le premier diagnostic) qui sont presque deux fois plus susceptibles de se produire en présence d'un carcinome lobulaire. En revanche, aucune association significative n'est trouvée avec les tumeurs primaires controlatérales détectées plus de 6 mois après le diagnostic.

**Sous-réseau chirurgie et traitement ultérieur.** Il est intéressant de noter que iMIIC met également en évidence le rôle central de la chirurgie dans la caractérisation précise des tumeurs primaires, Fig. 4a. Par exemple, iMIIC met en évidence un lien de causalité authentique quelque peu inattendu entre la chirurgie et l'histologie, qui reflète le fait que les types histologiques sont souvent affinés après la chirurgie par le pathologiste sur la base de la pièce opératoire. Ceci est cohérent avec une augmentation significative des types histologiques incluant des tissus spécifiques après l'intervention chirurgicale, comme le carcinome canalaire infiltrant mélangé à d'autres types de carcinome (+77 % après l'intervention chirurgicale), le carcinome canalaire infiltrant et le carcinome lobulaire (+48 %), carcinome canalaire infiltrant, sans autre précision (+7. 6 %), et une diminution correspondante des types histologiques plus génériques tels que le carcinome lobulaire, NOS (-11 %), le carcinome, NOS (-91 %), et l'adénocarcinome, NOS (-95 %). De même, iMIIC déduit à juste titre que la variable de stadification, AJCC7thN, est généralement basée sur le rapport pathologique après chirurgie, alors que le fait de ne pas effectuer de chirurgie (en raison de la présence de métastases à distance au moment du diagnostic ou de l'âge avancé du patient) conduit à une localisation beaucoup plus fréquente du quadrant mammaire non spécifié pour la tumeur primaire, Fig. 4a, *i.e.* 30,4 % "Breast NOS" quand la chirurgie n'est pas recommandée *contre* 11,1 % quand elle est effectuée, Fig. 4b.

De même, iMIIC met en évidence le rôle central de la chirurgie sur les décisions thérapeutiques concernant les traitements subséquents, tels que la reconstruction mammaire et la radiothérapie, Fig. 4a. Si la reconstruction mammaire nécessite effectivement une chirurgie du sein, Fig. 4c, iMIIC déduit également à juste titre que le type de chirurgie au site primaire (tumorectomie aussi connu sous le nom mastectomie partielle, ou mastectomie) dépend largement du choix personnel des patientes atteintes d'un cancer du sein au stade précoce entre les alternatives de conservation et de reconstruction du sein, Fig. 4a,d. De même, iMIIC déduit à juste titre que la radiothérapie est une "conséquence" fréquente de la chirurgie mammaire, Fig. 4a, *i.e.* 53 % *contre* 4% de radiothérapie, que la chirurgie soit pratiquée ou non, Fig. 4e, surtout après une tumorectomie (75 %) pour limiter le risque de rechute après une chirurgie de conservation du sein.

**Sous-réseau socio-économique.** Le réseau complet du cancer du sein sur Fig. 2a comprend quatre variables socio-économiques se rapportant au comté de résidence de chaque patient : le revenu familial médian, le revenu médian des ménages, l'indice du coût de la vie et la taille de la population rurale/urbaine de chaque comté. Ces quatre variables socio-économiques forment en fait un sous-graphe entièrement connecté. (*i.e.* une clique), indiquant leurs fortes interdépendances, et sont directement liées à un certain nombre d'autres variables, Fig. 5a. Il est intéressant de noter que le status vital n'est relié à cette clique de variables de comté que par le biais du revenu médian des ménages, ce qui est cohérent avec des rapports antérieurs sur l'association entre l'espérance de vie et les revenus [43]. En revanche, toutes les autres variables spécifiques aux patients liées à la clique du socio-économique (telles que le grade de la tumeur, la radiothérapie, la reconstruction mammaire, l'assurance) ont en fait au moins un lien avec l'indice du coût de la vie, ce qui souligne l'intégration du système de santé dans l'économie nationale. En particulier, il existe une association directe entre un coût de la vie plus élevé et un pronostic plus favorable du cancer du sein (*e.g.* moins de composants invasifs au moment du diagnostic). Cela est probablement dû à de meilleurs soins préventifs, notamment un accès plus facile aux centres de dépistage du cancer du sein et une couverture d'assurance plus complète. Cependant, il existe également de fortes disparités entre les comtés, comme le montrent les associations opposées entre l'assurance et la radiothérapie et le revenu familial médian *contre* l'indice du coût de la vie, Fig. 5a. Ces résultats intrigants peuvent être attribués au comté de Los Angeles (L.A.), qui représente environ 10 % de l'ensemble des données et qui présente un revenu familial médian inférieur à la moyenne (29-38 % de la fourchette centile) malgré un indice du coût de la vie supérieur à la moyenne (58-67 % de la fourchette centile), Fig. 5b. Cela a dû entraîner une charge financière exacerbée pour un nombre important des 39 089 patients atteintes d'un cancer du sein diagnostiqué dans le comté de Los Angeles entre 2010 et 2016. Bien que 18 % de ces patients aient bénéficié d'une assurance Medicaid (contre

10 % dans l'ensemble des données), beaucoup ont dû opter pour une assurance privée abordable mais limitée, comprenant des politiques de co-paiement importantes, ou même devenir non assurés, surtout avant l'application de l'Affordable Care Act en janvier 2014 (3,4 % de non assurés en 2013 contre 1,5 % en 2014). Par conséquent, de nombreux patients de L.A. semblent avoir renoncé à suivre des traitements coûteux. En particulier, seulement 32,6 % des patients ont subi une radiothérapie à Los Angeles, contre 50 % des patients dans tout le pays, à l'exception du comté de Los Angeles, Fig. 5c, qui ne peut s'expliquer qu'en partie par les différences départementales dans la sous-déclaration de la radiothérapie des patients externes à SEER [32, 33]. De plus, environ 7 % des patients de Los Angeles semblent même avoir abandonné leur thérapie ou avoir déménagé dans un autre comté non inclus dans la base de données SEER (contre 1,5 % à l'échelle nationale, à l'exclusion du comté de Los Angeles), d'après la distribution de la durée de suivi qui diminue rapidement à Los Angeles par rapport au reste de l'ensemble de données, Fig. 5d. Cela correspond à la fraction des patients ayant eu leur dernier contact médical moins d'un an après le diagnostic et plus d'un an avant la fin de cette étude en décembre 2016.

## Discussion

Les données sur les soins de santé à l'échelle nationale, comme les données SEER sur le cancer du sein analysé ici, sont particulièrement intéressantes d'un point de vue méthodologique: elles fournissent des ensembles de données de référence, qui peuvent aider à évaluer la fiabilité des méthodes de découverte causale sur des données réelles. En effet, la plupart des prédictions de cause à effet peuvent être validées ou rejetées, sur la base des connaissances des experts, de la pratique clinique ou d'éventuels biais de collecte et de sélection des données. En outre, l'interprétabilité des méthodes d'apprentissage automatique est particulièrement importante pour les applications sur les données cliniques, pour lesquelles les recommandations assistées par l'intelligence artificielle peuvent difficilement s'appuyer uniquement sur des classificateurs "boîte noire" et doivent pouvoir être expliquées en termes de raisonnements intelligibles aux praticiens médicaux [44]. Pourtant, au-delà des données cliniques, les méthodes de découverte causale ont le potentiel de devenir des approches d'apprentissage automatique essentielles pour interpréter diverses données d'observation dans un large éventail de domaines, pour lesquels des expériences de perturbation systématique ne sont pas disponibles pour des raisons pratiques, financières ou éthiques. En particulier, la découverte causale peut guider la recherche biologique en prédisant les effets causaux d'interventions spécifiques [45], tels que l'expression ou l'extinction des gènes, qui peuvent ensuite être étudiés par des expériences ciblées de siRNA, de knock-out génique ou d'édition basée sur CRISPR.

Dans le contexte de l'ensemble de données du SEER sur le cancer du sein, iMIIC met en évidence de nombreuses relations causales attendues, telles que la conséquence négative des métastases et l'effet protecteur du statut ER+ sur le décès dû au cancer du sein, ou le fait que l'année de naissance est la principale raison du décès dû à d'autres causes à la fin de l'étude. D'autre part, les effets de la couverture d'assurance ou de l'état matrimonial, qui ont été signalés comme réduisant le risque de décès dû au cancer du sein, s'avèrent être entièrement indirects et principalement médiés par les traitements (60-80 %), notamment, la chirurgie (>50 %). En effet, la chirurgie apparaît comme la pierre angulaire de la thérapie du cancer du sein en permettant d'abord d'affiner les types histologiques, puis de guider les décisions thérapeutiques en matière de radiothérapie, de reconstruction mammaire et enfin de prolonger les délais de survie des patientes. Pourtant, iMIIC déduit également à juste titre que le type de chirurgie (tumorectomie ou mastectomie) au niveau du site primaire dépend largement du choix personnel des patientes atteintes d'un cancer du sein au stade précoce entre les alternatives de conservation ou de reconstruction du sein. En revanche, d'autres traitements, comme la radiothérapie et la chimiothérapie, semblent avoir un impact moins décisif sur l'issue du cancer du sein, ce qui pourrait être dû en partie à la sous-déclaration de certaines informations sur les traitements dans la base de données SEER. [32, 33]. La radiothérapie semble même être une conséquence, et non une cause, du statut vital, ce qui suggère qu'un décès précoce dans les premiers mois après le diagnostic peut empêcher l'accès à la radiothérapie pour certains patients qui auraient pu autrement recevoir ce traitement, s'ils avaient vécu plus longtemps. Enfin, iMIIC retrouve des associations directes entre les variables socio-économiques du comté (telles que le revenu familial médian et l'indice du coût de la vie) et les variables spécifiques au patient (telles que le grade de la tumeur, la radiothérapie, la reconstruction mammaire, l'assurance), soulignant l'intégration du système de santé dans l'économie nationale. Si un coût de la vie plus élevé est en moyenne associé à un pronostic de cancer plus favorable, probablement en raison de l'amélioration des soins de santé préventifs et d'une couverture d'assurance plus complète, iMIIC révèle

également de grandes disparités entre les indices de revenu familial et de coût de la vie dans les différents comté. (*e.g.* pour le comté de L.A.), conduisant à une charge financière exacerbée, avec des patients renonçant alors à des traitements coûteux ou même abandonnant la thérapie.

En résumé, iMIIC est une méthode générale de découverte causale, qui permet de découvrir des relations directes et éventuellement causales ainsi que des effets indirects cohérents avec le réseau pour un large éventail de données biologiques et cliniques. Il est important de noter que iMIIC est entièrement non supervisé et ne nécessite pas d'hypothèses préalables ni de connaissances spécialisées. En particulier, iMIIC ajuste automatiquement les facteurs de confusion mesurés (sous la forme de contributions indirectes) et distingue les causes réelles des effets causaux putatifs et latents en excluant ou en mettant en évidence l'effet des facteurs de confusion non mesurés pour chaque arête causale (Encadré 1). Bien que iMIIC ne soit pas à l'abri d'éventuels biais de collecte et de sélection des données, qui peuvent affecter les données d'observation, il repose sur un cadre robuste de la théorie de l'information, ce qui le rend particulièrement fiable pour interpréter des types de données difficiles, comme les données hétérogènes comprenant une combinaison de variables continues et catégorielles intégrées à partir de différentes sources (*e.g.* des données cliniques, personnelles, socio-économiques, comme le montre le cas ici) ou des techniques expérimentales différentes (*e.g.* données transcriptomiques single-cell [8, 45, 46] et données d'imagerie [10]). En principe, iMIIC pourrait être appliqué à un large éventail d'autres domaines pour découvrir des relations causales et quantifier des contributions indirectes lorsque seules des données d'observation sont disponibles. Avec l'avènement d'ensembles de données pratiquement illimités dans de nombreux domaines de la science des données, les méthodes évolutives de découverte causale sont nécessaires et iMIIC peut répondre à ce besoin.

# References


1. Hulswit, M. *From cause to causation: A Peircean perspective* (Springer Science & Business Media, 2002).
2. Pearl, J. & Mackenzie, D. *The book of why: the new science of cause and effect* (Basic books, 2018).
3. Spirtes, P., Glymour, C. N., Scheines, R. & Heckerman, D. *Causation, prediction, and search* (MIT press, 2000).
4. Pearl, J. *Causality* (Cambridge university press, 2009).
5. Heckerman, D., Geiger, D. & Chickering, D. M. Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. *Mach. Learn.* **20,** 197–243 (1995).
6. Friedman, J., Hastie, T. & Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9,** 432–441 (2008).
7. Zhang, J. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artif. Intell.* **172,** 1873–1896 (2008).
8. Verny, L., Sella, N., Affeldt, S., Singh, P. P. & Isambert, H. Learning causal networks with latent variables from multivariate information in genomic data. *PLoS Comput. Biol.* **13,** e1005662 (2017).
9. Sella, N., Verny, L., Uguzzoni, G., Affeldt, S. & Isambert, H. MIIC online: a web server to reconstruct causal or non-causal networks from non-perturbative data. *Bioinformatics* **34,** 2311–2313 (2018).
10. Cabeli, V. *et al.* Learning clinical networks from medical records based on information estimates in mixed-type data. *PLOS Computational Biology* **16,** e1007866 (2020).
11. Howlader, N. *et al.* in *SEER Cancer Statistics Review* 1975–2016 (National Cancer Institute, 2018).
12. Peters, J., Mooij, J. M., Janzing, D. & Schölkopf, B. Causal Discovery with Continuous Additive Noise Models. *Journal of Machine Learning Research* **15,** 2009–2053 (2014).
13. Sackett, D. L. Bias in analytic research. *Journal of Chronic Diseases* **32,** 51–63 (1979).
14. Hernán, M. A., Hernández-Díaz, S. & Robins, J. M. A Structural Approach to Selection Bias. *Epidemiology* **15,** 615–625 (2004).

15. Prosperi, M. *et al.* Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence* **2,** 369–375 (2020).

16. Cover, T. M. & Thomas, J. A. *Elements of Information Theory* 2nd (Wiley, 2006).

17. Cabeli, V., Li, H., da Câmara Ribeiro-Dantas, M., Simon, F. & Isambert, H. *Reliable causal discovery based on mutual information supremum principle for finite datasets* in *why21 at 35rd Conference on Neural Information Processing Systems* (NeurIPS, 2021).

18. Li, H., Cabeli, V., Sella, N. & Isambert, H. *Constraint-based causal structure learning with consistent separating sets* in *33rd Conference on Neural Information Processing Systems* (NeurIPS, 2019).

19. Spirtes, P. & Glymour, C. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review* **9,** 62–72 (1991).

20. Kalisch, M., Mächler, M., Colombo, D., Maathuis, M. H. & Bühlmann, P. Causal inference using graphical models with the R package pcalg. *J. Stat. Softw.* **47,** 1–26 (2012).

21. Ramsey, J., Spirtes, P. & Zhang, J. *Adjacency-Faithfulness and Conservative Causal Inference* in *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence* (AUAI Press, 2006), 401–408.

22. Sedgewick, A. J. *et al.* Mixed graphical models for integrative causal analysis with application to chronic lung disease diagnosis and prognosis. *Bioinformatics* **35,** 1204–1212 (2018).

23. Tsagris, M., Borboudakis, G., Lagani, V. & Tsamardinos, I. Constraint-based causal discovery with mixed data. *International Journal of Data Science and Analytics* **6,** 19–30 (2018).

24. Harbeck, N. *et al.* Breast cancer. *Nature Reviews Disease Primers* **5,** 6 (2019).

25. Alaa, A. M., Gurdasani, D., Harris, A. L., Rashbass, J. & van der Schaar, M. Machine learning to guide the use of adjuvant therapies for breast cancer. *Nature Machine Intelligence* **3,** 716–726 (2021).

26. Lee, C. *et al.* Application of a novel machine learning framework for predicting non-metastatic prostate cancer-specific mortality in men using the Surveillance, Epidemiology, and End Results (SEER) database. *The Lancet Digital Health* **3,** e158–e165 (2021).

27. Welch, H. G., Prorok, P. C., O'Malley, A. J. & Kramer, B. S. Breast-Cancer Tumor Size, Overdiagnosis, and Mammography Screening Effectiveness. *New England Journal of Medicine* **375,** 1438–1447 (2016).

28. Leapman, M. S. *et al.* Mediators of Racial Disparity in the Use of Prostate Magnetic Resonance Imaging Among Patients With Prostate Cancer. *JAMA Oncology, Published online March 03* (2022).

29. Petito, L. C. *et al.* Estimates of Overall Survival in Patients With Cancer Receiving Different Treatment Regimens. *JAMA Network Open* **3,** e200452 (2020).

30. Nethery, R. C., Yang, Y., Brown, A. J. & Dominici, F. A causal inference framework for cancer cluster investigations using publicly available data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **183,** 1253–1272 (2020).

31. Wang, L. Mining causal relationships among clinical variables for cancer diagnosis based on Bayesian analysis. *BioData Mining* **8,** 13 (2015).

32. Park, H. S., Lloyd, S., Decker, R. H., Wilson, L. D. & Yu, J. B. Limitations and Biases of the Surveillance, Epidemiology, and End Results Database. *Current Problems in Cancer* **36,** 216–224 (2012).

33. Jagsi, R. *et al.* Underascertainment of radiotherapy receipt in Surveillance, Epidemiology, and End Results registry data. *Cancer* **118,** 333–341 (2011).

34. Chen, S.-Y. *et al.* Timing of Chemotherapy and Radiotherapy Following Breast-Conserving Surgery for Early-Stage Breast Cancer: A Retrospective Analysis. *Frontiers in Oncology* **10,** 571390 (2020).

35. Anderson, J. R., Cain, K. C. & Gelber, R. D. Analysis of survival by tumor response. *Journal of Clinical Oncology* **1,** 710–719 (1983).

36. Hernán, M. A. & Robins, J. M. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available: Table 1. *American Journal of Epidemiology* **183,** 758–764 (2016).

37. Han, X., Yabroff, K. R., Ward, E., Brawley, O. W. & Jemal, A. Comparison of Insurance Status and Diagnosis Stage Among Patients With Newly Diagnosed Cancer Before vs After Implementation of the Patient Protection and Affordable Care Act. *JAMA Oncology* **4,** 1713 (2018).

38. Ermer, T. *et al.* Understanding the Implications of Medicaid Expansion for Cancer Care in the US. *JAMA Oncology* **8,** 139 (2022).

39. Hinyard, L., Wirth, L. S., Clancy, J. M. & Schwartz, T. The effect of marital status on breast cancer-related outcomes in women under 65: A SEER database analysis. *The Breast* **32,** 13–17 (2017).

40. Zhai, Z. *et al.* Effects of marital status on breast cancer survival by age, race, and hormone receptor status: A population-based Study. *Cancer medicine* **8,** 4906–4917 (2019).

41. Bonéy-Montoya, J., Ziegler, Y. S., Curtis, C. D., Montoya, J. A. & Nardulli, A. M. Long-range transcriptional control of progesterone receptor gene expression. *Mol Endocrinol* **24,** 346–358 (2010).

42. Fisher, C. *et al.* Histopathology of breast cancer in relation to age. *British journal of cancer* **75,** 593–596 (1997).

43. Chetty, R. *et al.* The Association Between Income and Life Expectancy in the United States, 2001-2014. *JAMA* **315,** 1750 (2016).

44. Binder, A. *et al.* Morphological and molecular breast cancer profiling through explainable machine learning. *Nature Machine Intelligence* **3,** 355–366 (Mar. 2021).

45. Desterke, C. *et al.* Inferring Gene Networks in Bone Marrow Hematopoietic Stem Cell-Supporting Stromal Niche Populations. *iScience* **23,** 101222 (2020).

46. Affeldt, S., Verny, L. & Isambert, H. 3off2: A network reconstruction algorithm based on 2-point and 3-point information statistics. *BMC Bioinformatics* **17,** 12 (2016).

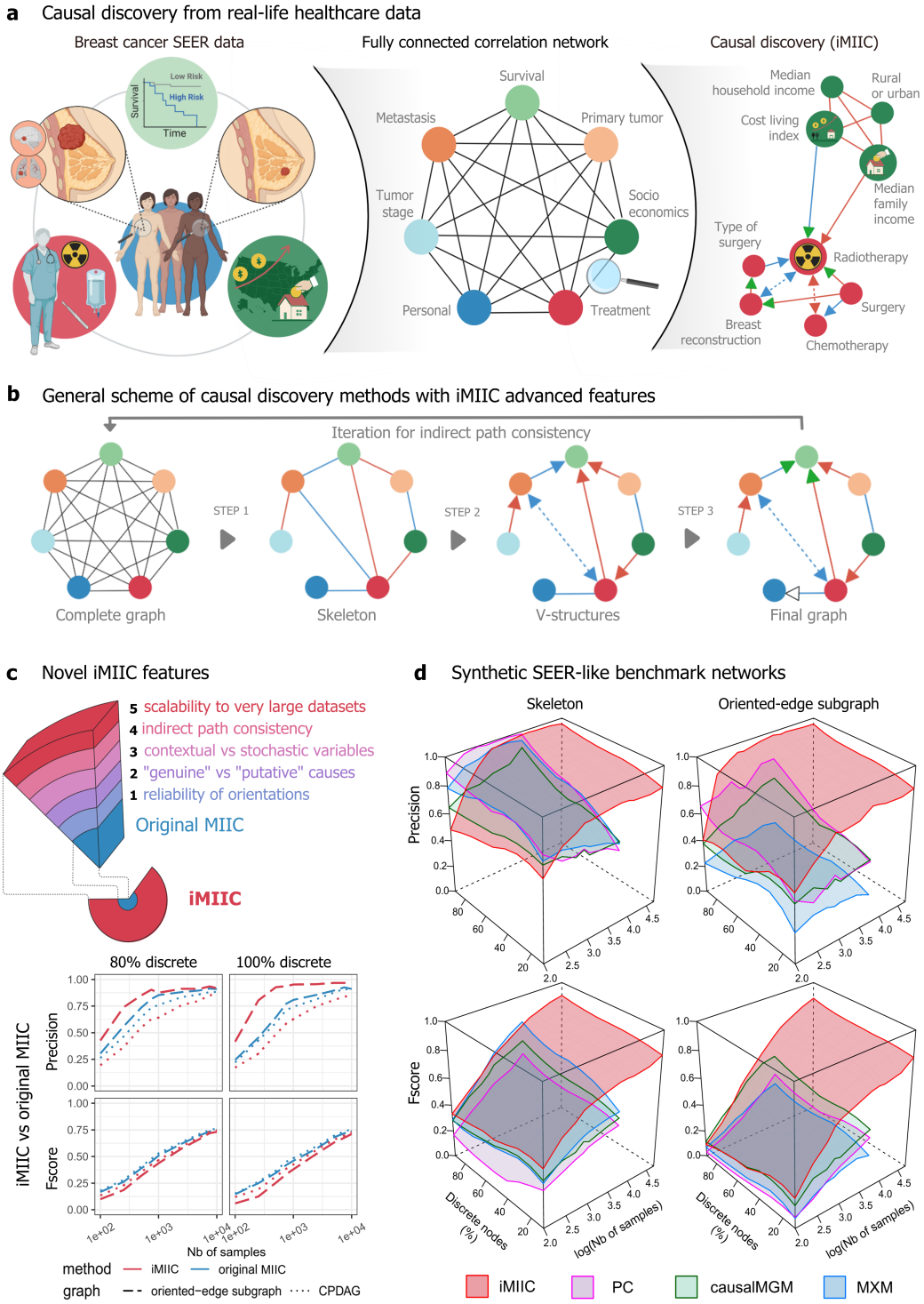
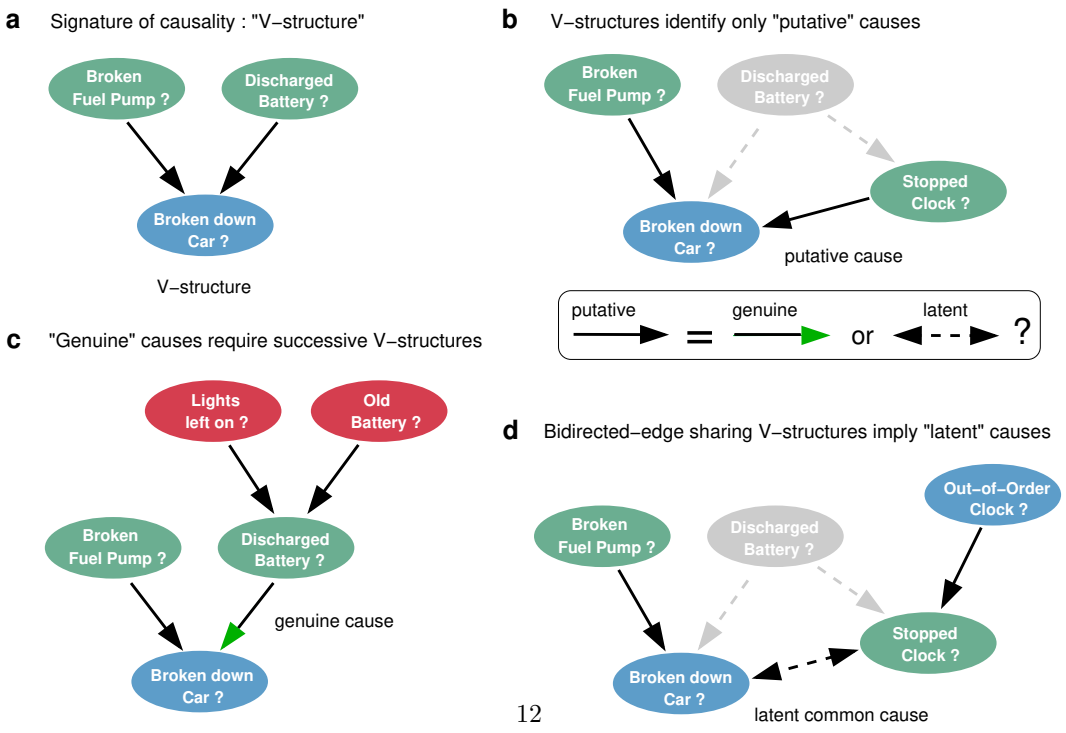



Figure 1: **Découverte causale à partir de données de santé réelles à l'aide de méthodes basées sur les contraintes. (a)** La base de données SEER comprend 407 791 dossiers médicaux de patients atteints d'un cancer du sein diagnostiqué entre 2010 et 2016. La découverte causale vise à découvrir les relations de cause à effet à travers de tels ensembles de données globalement corrélés. **(b)** Schéma général des méthodes basées sur les contraintes (y compris les nouvelles fonctionnalités avancées d'iMIIC) : Étape 1, suppression des arêtes inutiles (avec garantie de la cohérence des chemins indirects) ; Étape 2, orientation des 'structures en V' (les orientations fiables étant représentées par des flèches simples et les causes communes latentes par des arêtes bidirectionnelles); Étape 3, propagation de l'orientation indiquée par la pointe de flèche blanche (et distinction entre les causes 'putatives' et 'authentiques', pointes de flèche vertes). **(c)** Nouvelles fonctionnalités avancées de iMIIC et comparaison avec MIIC original. **(d)** Réseaux de référence synthétiques de type SEER avec différentes proportions de variables discrètes. Créé avec BioRender.com

## Encadré 1. Principes de découverte de causes à partir de données d'observation : causes putatives, authentiques et latentes.

Nous exposons ici les principes permettant de découvrir les relations de cause à effet dans un ensemble de données purement observationnelles et de distinguer les causes "authentiques" des causes "putatives" et des causes "latentes". Le raisonnement est illustré par l'exemple fictif, intuitif et causal, d'un ensemble de données imaginaire de vieilles voitures. (**a**) La signature de la causalité dans ces ensembles de données d'observation correspond à des sous-graphes de 3 variables appelés "structure en V" impliquant deux causes possibles *indépendantes* et donc *non connectées*, "Broken fuel pump?" et "Discharged battery?", et un effet résultant, "Broken down car?". Les orientations convergentes de cette structure en V vers sa variable centrale, "Broken down car?", proviennent du fait que ces deux arêtes ne peuvent pas être non orientés, et qu'elles ne peuvent pas non plus pointer vers "Broken fuel pump?" ou "Discharged battery?", car ces modèles graphiques alternatifs impliqueraient des corrélations contredisant l'indépendance entre "Broken fuel pump?" et "Discharged battery?". Alternativement, les relations causales peuvent parfois être découvertes entre deux variables seulement, sous l'hypothèse spécifique de modèles de bruit additif continu [12]. Cependant, dans le cas général, la découverte causale nécessite au moins trois variables et souvent plus, car l'indépendance entre les causes possibles dans une structure en v est fréquemment conditionnée par une ou plusieurs autres variables, non considérées ici, définissant un ensemble séparateur.
A l'inverse, le conditionnement sur la pointe d'une structure en V, ici "Broken down car?", induit des associations erronées entre ses causes possibles indépendantes [3, 4]. De même, la sélection d'un ensemble de données présentant des valeurs spécifiques pour cette variable de référence entraîne des associations erronées dues à un biais de sélection dans l'ensemble de données [13–15], comme une anti-corrélation apparente entre différentes causes possibles, "Broken fuel pump?" et "Discharged battery?", si seulement "Broken down car? = yes" sont sélectionnés. (**b**) Cependant, les structures en V restent en fait causalement ambiguës [4] car ils n'identifient que des causes "putatives", qui peut être soit une cause "authentique", affichée par une flèche verte, soit la présence de facteurs de confusion non mesurés, *i.e.* des causes communes latentes non observées dans l'ensemble de données et représentées par une arête bidirectionnelle. Par exemple, la variable "Clock stopped?", fréquemment utilisé comme un substitut de "Discharged battery?", forme également une structure en V similaire avec "Broken fuel pump?"; Pourtant, il est bien connu que "Clock stopped?" ne peut pas être une véritable cause de "Broken down car?", comme le fait de trafiquer l'horloge d'une voiture ne peut pas réellement provoquer une panne de voiture. (**c**) En l'absence de connaissances de base et d'intervention directe sur les variables, montrer que "Discharged battery?" est en effet une véritable cause de "Broken down car?" nécessite d'exclure la possibilité d'une cause commune non observée (*i.e.* un facteur de confusion non mesuré) entre "Discharged battery" et "Broken down car?". Pour ce faire, il faut trouver une autre structure en V en amont de "Discharged battery?" (*e.g.* "Lights left on?"→"Discharged battery?"←"Old battery?") ou avoir une connaissance préalable d'une cause en amont (putative) et de montrer que l'effet de la "Broken down car?" est entièrement *indirect* et médiée (au moins en partie) par la variable intermédiaire "Discharged battery?". Cela nécessite de trouver une indépendance conditionnelle entre une variable amont et "Broken down car?" conditionné à un ensemble séparateur, qui comprend la variable intermédiaire "Discharged battery". (**d**) Inversement, exclure une cause putative en tant que cause authentique exige de démontrer que la relation provient effectivement d'une cause commune non observée en trouvant une quatrième variable (*e.g.* "Out-of-order clock?") définissant une autre structure en V, induisant une arête bidirectionnelle entre "Broken down car?" et "Clock stopped?" avec la structure en V en (b).

La méthode iMIIC distingue les arêtes causales authentiques des arêtes causales putatives, ainsi que les arêtes non dirigées et bidirigées, en évaluant des probabilités d'orientation de tête ou de queue distinctes à chaque extrémité de l'arête (cf. Résultats et méthodes).



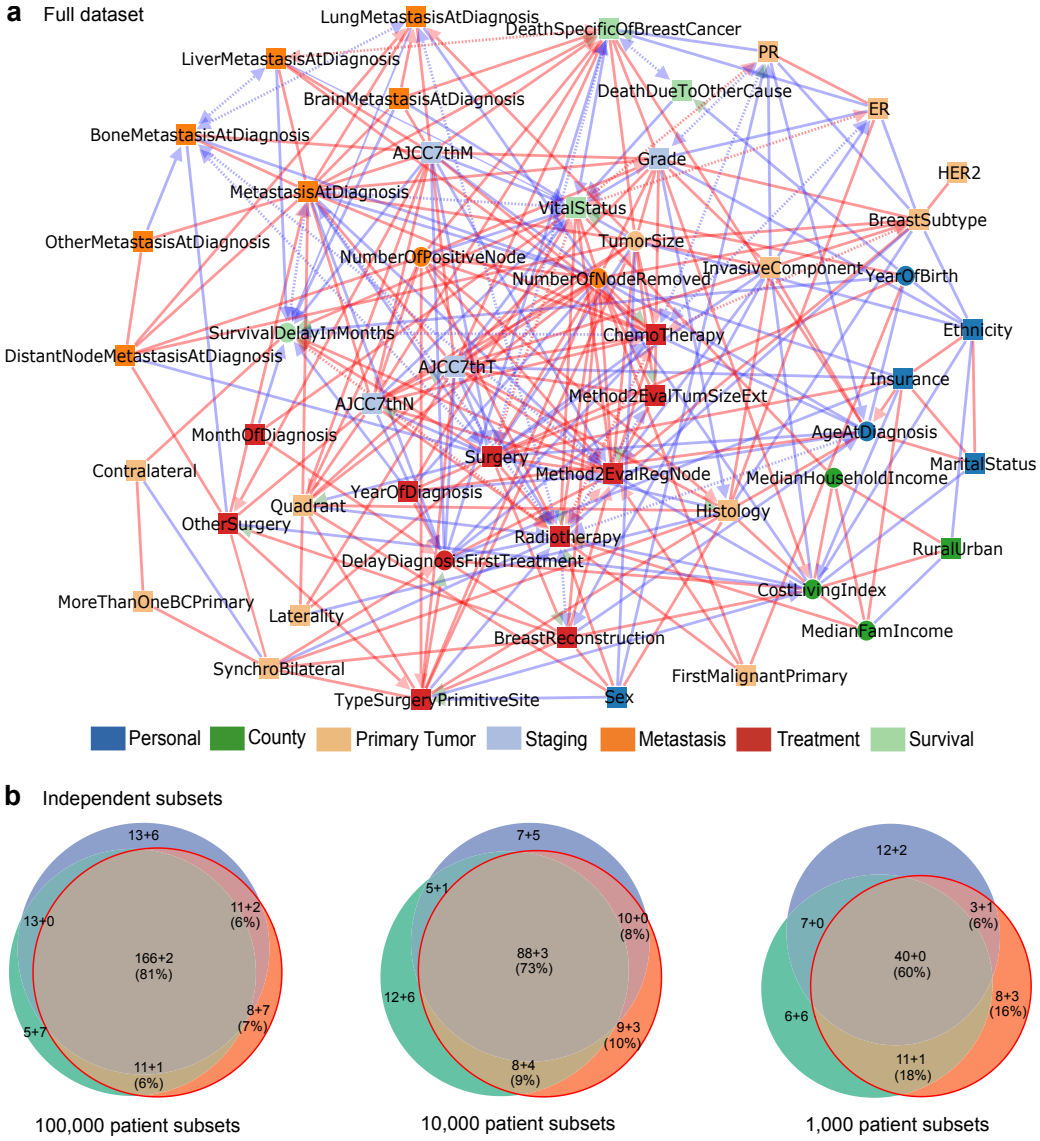



Figure 2: **Réseaux de cancer du sein SEER déduits par iMIIC**. **(a)** Le réseau de 51 nœuds déduit par iMIIC à partir de l'ensemble de données SEER comprenant 396 179 patientes atteintes d'un cancer du sein diagnostiqué entre 2010 et 2016. Ce réseau squeletto cohérent contient 280 arêtes et inclut 2 variables contextuelles, le sexe et l'année de naissance. Le réseau oriento cohérent correspondant contient 340 arêtes. **(b)** Comparaison de réseaux déduits de trois sous-échantillonnages indépendants de même taille de 100 000, 10 000 ou 1 000 sous-ensembles de patients (de gauche à droite). Le nombre d'arêtes partagées (indépendamment des orientations) dans les diagrammes d'Euler est donné sous forme de somme. $a+b$ où $a$ (resp. $b$) correspond au nombre d'arêtes incluses dans (resp. absentes de) le réseau complet de l'ensemble de données en (a). Les pourcentages se réfèrent au sous-ensemble de réseaux ayant le nombre total médian d'arêtes (cercle rouge).

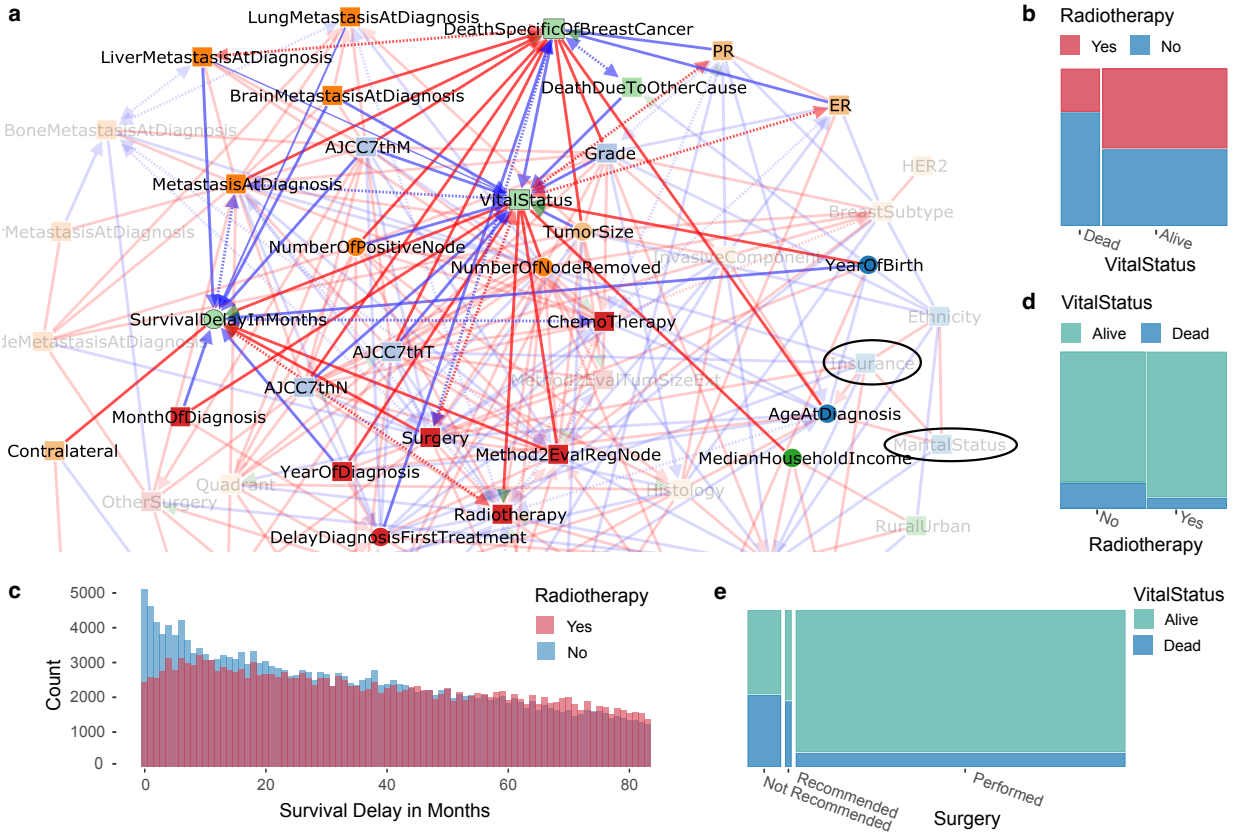



Figure 3: **Sous-réseau de survie déduit par iMIIC à partir de la base de données SEER sur le cancer du sein**. **(a)** Sous-réseau mettant en évidence les relations directes avec les variables de survie (VitalStatus, DeathSpecificOfBreastCancer, DeathDueToOtherCause, SurvivalDelayInMonths). L'absence de liens directs avec d'autres variables (telles que l'assurance et l'état matrimonial mis en évidence dans le réseau) peut être interprétée en termes de contributions de chemins indirects conformes au squelette du réseau. **(b)** Distribution conjointe de la radiothérapie et status vital mettant en évidence la relation causale contre-intuitive entre eux, voir le texte. **(c)** Histogramme du délai de survie en mois pour les patients ayant reçu ou non une radiothérapie. Chaque case représente un mois. Le pic bleu précoce suggère qu'un certain nombre de patients sont décédés dans les 3 à 6 mois suivant le diagnostic, donc avant d'avoir pu recevoir une radiothérapie, en accord avec la direction causale prédite dans (a). Il en résulte une surestimation du bénéfice apparent de la radiothérapie dans les cas (d), see main text. **(d)** Distribution conjointe du status vital et de la radiothérapie. **(e)** Distribution conjointe du status vital et de la chirurgie.

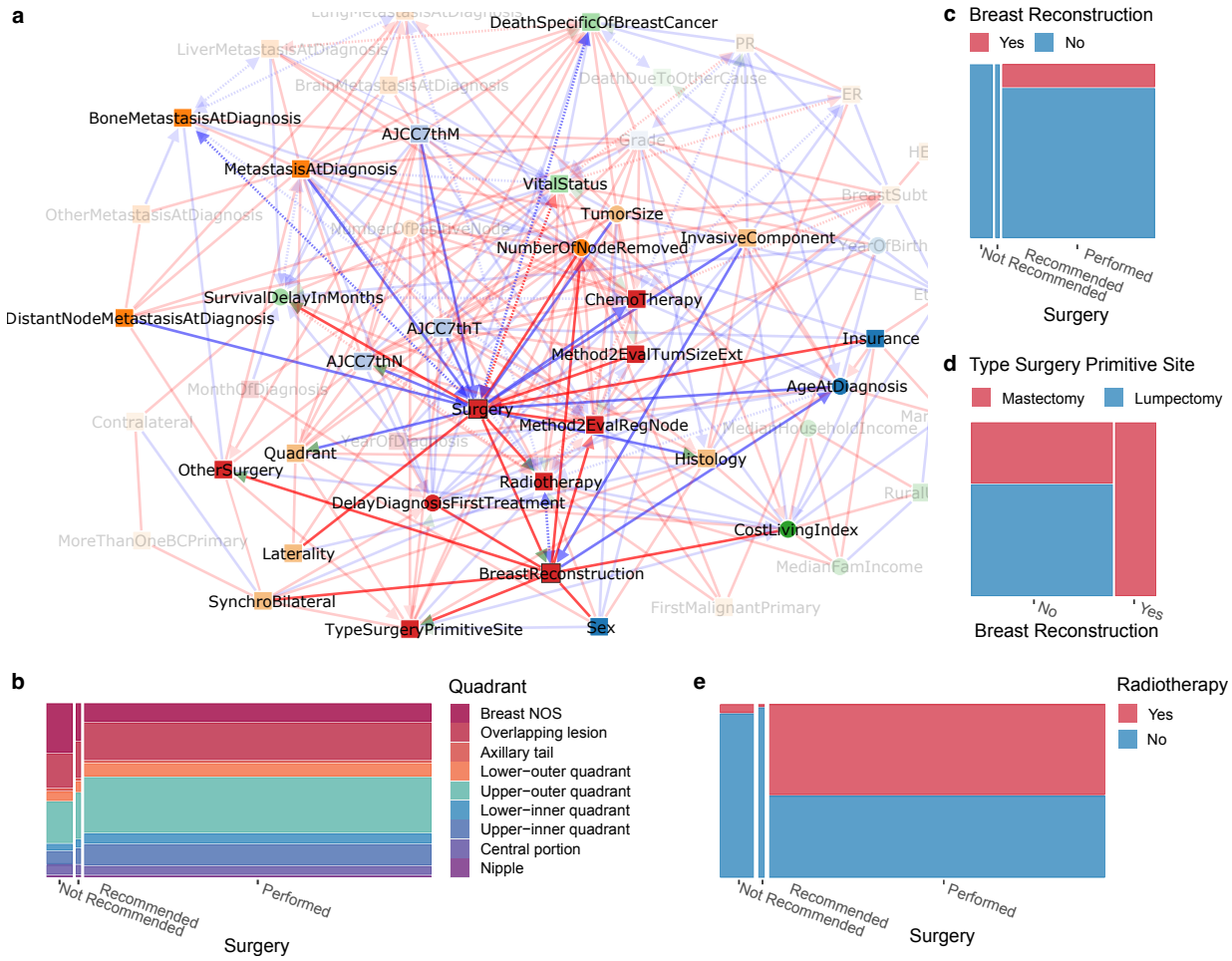



Figure 4: **Sous-réseau de la chirurgie et des traitements ultérieurs déduit par iMIIC à partir de l'ensemble de données SEER sur le cancer du sein**. **(a)** Sous-réseau mettant en évidence les relations directes avec la chirurgie et la reconstruction mammaire. **(b)** Distribution conjointe de Quadrant et de Surgery. **(c)** Distribution conjointe de la reconstruction et de la chirurgie du sein. **(d)** Distribution conjointe de la chirurgie de type site primitif et de la reconstruction mammaire. **(e)** Distribution conjointe de la radiothérapie et de la chirurgie. Voir le texte principal pour l'interprétation causale du rôle de la chirurgie dans l'affinement de la caractérisation de la tumeur primaire et des décisions thérapeutiques ultérieures, y compris le choix personnel des patients.

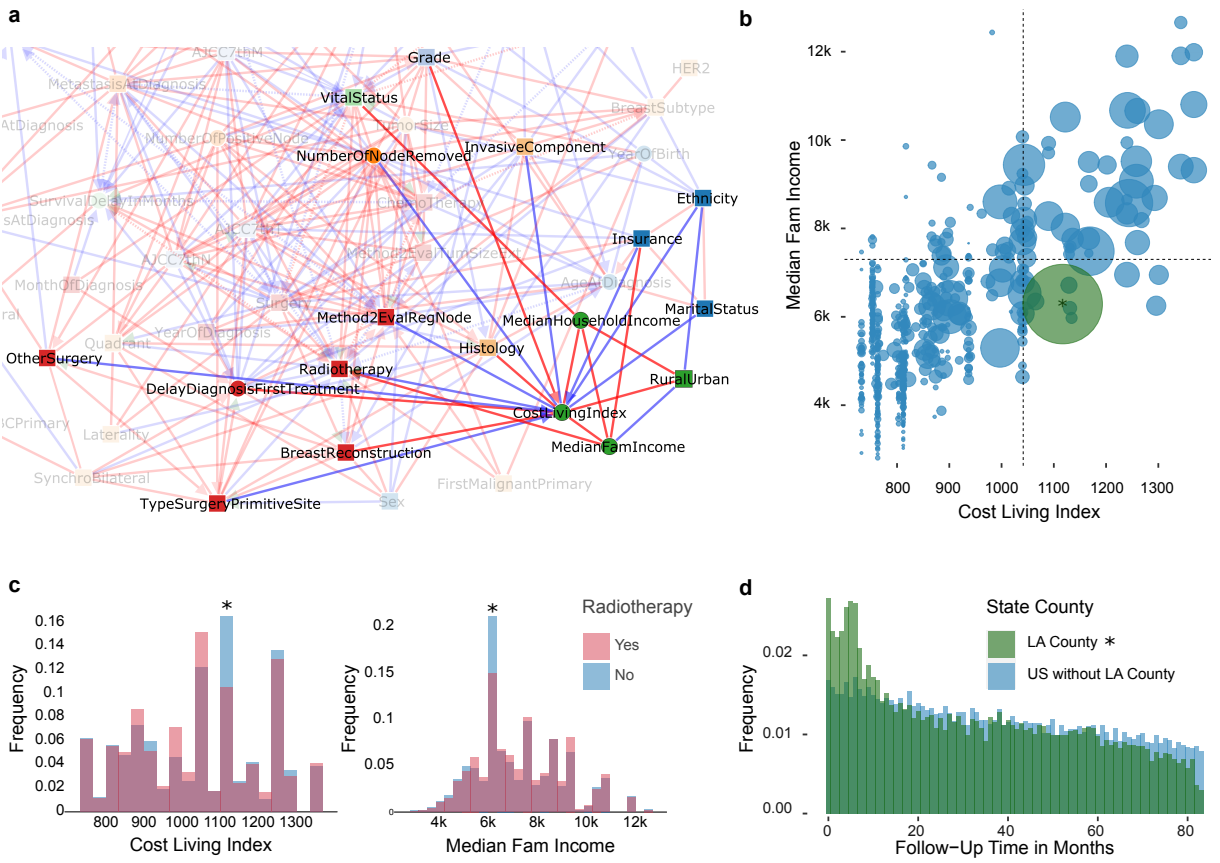



Figure 5: **Sous-réseau socio-économique déduit par iMIIC à partir de la base de données SEER sur le cancer du sein**. **(a)** Sous-réseau mettant en évidence les relations directes avec les variables socio-économiques du comté (CostLivingIndex, MedianFamIncome, MedianHouseholdIncome, et RuralUrban). **(b)** Graphique à bulles de la distribution conjointe du revenu familial médian et de l'indice du coût de la vie. L'aire des bulles représente le nombre de patients dans ce comté. Les lignes pointillées correspondent à l'indice moyen du coût de la vie et au revenu familial médian moyen. La bulle verte avec un astérisque correspond au comté de Los Angeles (L.A.) qui représente 10 % de l'ensemble des données. **(c)** Histogrammes de l'indice du coût de la vie et du revenu familial médian regroupés par radiothérapie. Les barres avec un astérisque correspondent au comté de Los Angeles. **(d)** Histogramme du temps de suivi en mois pour les patients de L.A. et pour tous les autres comtés américains inclus dans le SEER.

# References


Séverine Affeldt and Hervé Isambert. Robust reconstruction of causal graphical models based on conditional 2-point and 3-point information. In *ACI@ UAI*, pages 1–29, 2015.

Séverine Affeldt and Hervé Isambert. 3off2: A network reconstruction algorithm based on 2-point and 3-point information statistics. In *BMC bioinformatics*, volume 17, pages 149–165. BioMed Central, 2016.

Aristotle Aristotle. *Posterior analytics*, volume 1. Clarendon Press Oxford, UK, 1994.

José M Bernardo and Adrian FM Smith. *Bayesian theory*, volume 405. John Wiley & Sons, 2009.

Vincent Cabeli, Louis Verny, Nadir Sella, Guido Uguzzoni, Marc Verny, and Hervé Isambert. Learning clinical networks from medical records based on information estimates in mixed-type data. *PLoS computational biology*, 16(5):e1007866, 2020.

Sean M Carroll and Jennifer Chen. Spontaneous inflation and the origin of the arrow of time. *arXiv preprint hep-th/0410270*, 2004.

Clive R Charig, David R Webb, Stephen Richard Payne, and John E Wickham. Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy. *Br Med J (Clin Res Ed)*, 292(6524): 879–882, 1986.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

Carlos Cinelli, Jeremy Ferwerda, and Chad Hazlett. sensemakr: Sensitivity analysis tools for ols in r and stata. *Available at SSRN 3588978*, 2020a.

Carlos Cinelli, Andrew Forney, and Judea Pearl. A crash course in good and bad controls. *Available at SSRN 3689437*, 2020b.

Robin G Collingwood. On the so-called idea of causation. In *Proceedings of the Aristotelian society*, volume 38, pages 85–112. JSTOR, 1937.

Diego Colombo, Marloes H Maathuis, Markus Kalisch, and Thomas S Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, pages 294–321, 2012.

Diego Colombo, Marloes H Maathuis, et al. Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.*, 15(1):3741–3782, 2014.

Jerome Cornfield, William Haenszel, E Cuyler Hammond, Abraham M Lilienfeld, Michael B Shimkin, and Ernst L Wynder. Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer institute*, 22 (1):173–203, 1959.

Scott Cunningham. *Causal Inference: The Mixtape*. Yale University Press, 2021.

Neil M Davies, Michael V Holmes, and George Davey Smith. Reading mendelian randomisation studies: a guide, glossary, and checklist for clinicians. *Bmj*, 362, 2018.

Connor A Emdin, Amit V Khera, and Sekar Kathiresan. Mendelian randomization. *Jama*, 318(19):1925–1926, 2017.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

Joel B Greenhouse. Commentary: Cornfield, epidemiology and causality. *International journal of epidemiology*, 38(5):1199–1201, 2009.

MA Hernán and JM Robins. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC, 2020.

Sonia Hernández-Díaz, Enrique F Schisterman, and Miguel A Hernán. The birth weight "paradox" uncovered? *American journal of epidemiology*, 164(11):1115–1120, 2006.

Thomas Hobbes. *De Corpore (Concerning Body)*. The English Works Vol. I. Trans. W. Molesworth. London: Scientia Aalen, 1839.

Menno Hulswit. A short history of causation. *SEED Journal (Semiotics, Evolution, Energy, and Development)*, 4(3):16–42, 2004.

Immanuel Kant. 1998. critique of pure reason. *Trans. Paul Guyer and Allen Wood. Cambridge, UK: Cambridge University Press*, 1781.

JinHyung Kim and Judea Pearl. A computational model for causal and diagnostic reasoning in inference systems. In *International Joint Conference on Artificial Intelligence*, pages 0–0, 1983.

Jonathan Lamb. Scurvy - the disease of discovery. In *Scurvy - The Disease of Discovery*. Princeton University Press, 2016.

David Lewis. Causation. *Journal of Philosophy*, pages 556–567, 1973.

Honghao Li, Vincent Cabeli, Nadir Sella, and Hervé Isambert. Constraint-based causal structure learning with consistent separating sets. *Advances in Neural Information Processing Systems*, 32, 2019.

James Lind. A treatise of the scurvy. *Three Parts. Containing an Inquiry into the Nature, Causes and Cure, of that Disease. Together with a Critical and Chronological View of what has been Published on the Subject*, 1753.

John Locke. *An essay concerning human understanding*. Kay & Troutman, 1847.

AA Long. Stoic studies. berkeley, 1996.

Ida EH Madsen, Hermann Burr, Finn Diderichsen, Jan H Pejtersen, Marianne Borritz, Jakob B Bjorner, and Reiner Rugulies. Work-related violence and incident use of psychotropics. *American journal of epidemiology*, 174(12):1354–1362, 2011.

O.S. Miettinen. *Epidemiological Research: Terms and Concepts*. SpringerLink : Bücher. Springer Netherlands, 2011. ISBN 9789400711716. URL https://books.google.fr/books?id=GT-XHKhrrwgC.

John Stuart Mill. *A System of Logic*. Eighth edition. New York: Harper and Brothers., 1874.

RP Miller. *René Descartes: Principles of Philosophy: Translated, with Explanatory Notes*, volume 24. Springer Science & Business Media, 1984.

Leonard Mlodinow. *The drunkard's walk: How randomness rules our lives*. Vintage, 2009.

Jacob M Montgomery, Brendan Nyhan, and Michelle Torres. How conditioning on posttreatment variables can ruin your experiment and what to do about it. *American Journal of Political Science*, 62(3):760–775, 2018.

Isaac Newton. *The mathematical principles of natural philosophy (Philosophiae Naturalis Principia)*. 1687.

Judea Pearl. *Causality: Models, reasoning and inference*. Cambridge, UK: CambridgeUniversityPress, 2000.

Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.

Steven Piantadosi. *Clinical trials: a methodologic perspective*. John Wiley & Sons, 2017.

Plato Plato. *Timaeus*. BoD–Books on Demand, 2019.

Robert N Proctor. The history of the discovery of the cigarette–lung cancer link: evidentiary traditions, corporate denial, global toll. *Tobacco control*, 21(2):87–91, 2012.

Joseph Ramsey, Jiji Zhang, and Peter L Spirtes. Adjacency-faithfulness and conservative causal inference. *arXiv preprint arXiv:1206.6843*, 2012.

Hans Reichenbach. *The direction of time*, volume 65. Univ of California Press, 1956.

David L Sackett. Bias in analytic research. In *The case-control study consensus and controversy*, pages 51–63. Elsevier, 1979.

Andrew J Sedgewick, Ivy Shi, Rory M Donovan, and Panayiotis V Benos. Learning mixed graphical models with separate sparsity parameters and stability-based model selection. *BMC bioinformatics*, 17(5):307–318, 2016.

Andrew J Sedgewick, Kristina Buschur, Ivy Shi, Joseph D Ramsey, Vineet K Raghu, Dimitris V Manatakis, Yingze Zhang, Jessica Bon, Divay Chandra, Chad Karoleski, et al. Mixed graphical models for integrative causal analysis with application to chronic lung disease diagnosis and prognosis. *Bioinformatics*, 35(7):1204–1212, 2019.

Nadir Sella, Louis Verny, Guido Uguzzoni, Séverine Affeldt, and Hervé Isambert. Miic online: a web server to reconstruct causal or non-causal networks from non-perturbative data. *Bioinformatics*, 34(13):2311–2313, 2018.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvärinen, Yoshinobu Kawahara, Takashi Washio, Patrik O Hoyer, and Kenneth Bollen. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *The Journal of Machine Learning Research*, 12:1225–1248, 2011.

Edward H Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241, 1951.

Baruch Spinoza. 1677/1949 ethics. *Trans J Gutman New York Hafner Publishing Company*.

Peter Spirtes, Christopher Meek, and Thomas Richardson. An algorithm for causal inference in the presence of latent variables and selection bias. *Computation, causation, and discovery*, 21:211–252, 1999.

Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

Graham Sutton. Putrid gums and 'dead men's cloaths': James lind aboard the salisbury. *Journal of the Royal Society of Medicine*, 96(12):605–608, 2003.

Michail Tsagris, Giorgos Borboudakis, Vincenzo Lagani, and Ioannis Tsamardinos. Constraint-based causal discovery with mixed data. *International journal of data science and analytics*, 6(1):19–30, 2018.

Louis Verny, Nadir Sella, Séverine Affeldt, Param Priya Singh, and Hervé Isambert. Learning causal networks with latent variables from multivariate information in genomic data. *PLoS computational biology*, 13(10):e1005662, 2017.

Auguste Marie Philippa von Bayern, Samara Danel, AMI Auersperg, Berenika Mioduszewska, and A Kacelnik. Compound tool construction by new caledonian crows. *Scientific reports*, 8(1):1–8, 2018.

Kenneth E Warner, Donald W Sexton, Brenda W Gillespie, David T Levy, and Frank J Chaloupka. Impact of tobacco control on adult per capita cigarette consumption in the united states. *American Journal of Public Health*, 104(1):83–89, 2014.

Joan L Warren, Carrie N Klabunde, Deborah Schrag, Peter B Bach, and Gerald F Riley. Overview of the seer-medicare data: content, research applications, and generalizability to the united states elderly population. *Medical care*, pages IV3–IV18, 2002.

Daniel Westreich and Sander Greenland. The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. *American journal of epidemiology*, 177(4):292–298, 2013.

C. J. Wink, K. Woensdregt, G. A.P. Nieuwenhuijzen, M. J.C. Van Der Sangen, S. Hutschemaekers, J. A. Roukema, V. C.G. Tjan-Heijnen, and A. C. Voogd. Hormone treatment without surgery for patients aged 75 years or older with operable breast cancer. *Annals of Surgical Oncology*, 19(4):1185–1191, apr 2012. ISSN 10689265. doi: 10.1245/S10434-011-2070-Z.

J Yerushalmy. The relationship of parents' cigarette smoking to outcome of pregnancy—implications as to the problem of inferring causation from observed associations. *American Journal of Epidemiology*, 93:443–56, 1971.