Supplementary Material

Appendix A Background on Causality

A.1 Representing Cause-Effect Relationships with Graphs

A graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ is a pair of two sets, a set nodes $\mathbf{V}$ and a set of edges $\mathbf{E} \subseteq \mathbf{V} \times \mathbf{V}$. Two nodes X and Y are *adjacent* if an edge connects them. Edges can be directed, $X \to Y$, or undirected, $X - Y$. A *directed graph* is a graph in which every edge is directed. If $X \to Y$ then X is a *parent* of Y and Y is a *child* of X. A sequence of nodes $X_1, \ldots, X_k$ form a *path* if X_i and X_{i+1} are adjacent for every $i \in [1, k)$. A path is *directed* if $X_i \to X_{i+1}$ for at least one i [30]. X is an *ancestor* of Y, and Y is a *descendant* of X, if there exists a directed path from X to Y. We use $An(X)$ to indicate X's ancestors, and $De(X)$ to indicate X's descendants. A *cycle* is a directed path where $X_1 = X_k$. A graph is *acyclic* when there are no cycles. A DAG is a directed and acyclic graph.

Definition 1 (Causal graph) A *causal graph* (CG) [4] is a DAG $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ where each node $X_i \in \mathbf{V}$ represents a random variable. It is determined by a function f_i and a set of *exogenous* variables $\mathbf{U}_X$ so that:

$$X_i \coloneqq f_i(Pa(X_i), \mathbf{U}_X) \tag{A1}$$

with $Pa(X_i)$ being the parents of X_i.

The exogenous variables indicate the conditions of the world outside the system and are influenced by external factors. Typically, these variables represent the individuals involved in the phenomenon, belonging to the population under study [4]. Hereafter, we refer to the parents $Pa(X_i)$, including the variables $\mathbf{U}_X$.

A causal graph describes how a system under study works: for each edge $X \to Y$, the node X is the cause of Y and Y the effect of X [41]. *Observational data* are collected without intervening on causal mechanisms [41]. The causal networks framework [4, 40] allows us to relate a causal graph to the data distribution.

Definition 2 (Causal network) A *causal network* (CN) is a tuple $(\mathcal{G}, P)$ where $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ is a causal graph and P is a joint probability distribution over $\mathbf{V}$ that factorizes into local distributions accordingly to $\mathcal{G}$:

$$P(\mathbf{V}) = \prod_{X_i \in \mathbf{V}} P(X_i | Pa(X_i)) \tag{A2}$$

The condition in Equation (A2) makes the CN a *Bayesian network*, where edges also have a causal interpretation [30, 43]. Hence, CNs provide a causal and a probabilistic explanation of the system's behavior.

Definition 3 (d-separation) Given a DAG $\mathcal{G} = (\mathbf{V}, \mathbf{E})$, a path π between two nodes X and Y is *d-separated* [42] by a set $\mathbf{Z} \subseteq \mathbf{V} \setminus \{X, Y\}$ if:

- π contains a *chain* of nodes $X_{i-1} \to X_i \to X_{i+1}$ or a *fork* $X_{i-1} \leftarrow X_i \to X_{i+1}$ such that $X_i \in \mathbf{Z}$, or
- π contains a *collider* $X_{i-1} \to X_i \leftarrow X_{i+1}$ such that $X_i \notin \mathbf{Z}$, and no descendant of X_i are in $\mathbf{Z}$.

Two disjoint sets of nodes $\mathbf{X}$ and $\mathbf{Y}$ are d-separated by $\mathbf{Z} \subseteq \mathbf{V} \setminus \{\mathbf{X}, \mathbf{Y}\}$ if all pairs $(X, Y) \in \mathbf{X} \times \mathbf{Y}$ are d-separated by $\mathbf{Z}$. In that case, we write $\mathbf{X} \perp\!\!\!\perp_{\mathcal{G}} \mathbf{Y} \mid \mathbf{Z}$.

We denote by $\mathbf{X} \perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{Z}$ the conditional independence between the sets of random variables $\mathbf{X}$ and $\mathbf{Y}$ given the set $\mathbf{Z}$, i.e., whenever $P(\mathbf{X} \mid \mathbf{Y}, \mathbf{Z}) = P(\mathbf{X} \mid \mathbf{Z})$. The condition in Equation (A2) is equivalent to the following *Markov property* [30, 31]:

$$\mathbf{X} \perp\!\!\!\perp_{\mathcal{G}} \mathbf{Y} \mid \mathbf{Z} \;\Rightarrow\; \mathbf{X} \perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{Z} \tag{A3}$$

The Markov property allows us to read conditional independencies from the graph. However, many DAGs can encode the same set of d-separations. The DAGs that entail the same independencies of a given CG $\mathcal{G}$ belong to the same *Markov equivalence class* (MEC) of $\mathcal{G}$, denoted by $[\mathcal{G}]$ [43]. A MEC can be uniquely represented by a *completed partially DAG* (CPDAG), namely an acyclic graph containing directed and undirected edges, in which an edge is directed iff it is directed in all DAGs belonging to the MEC.

A.2 The Problem of Causal Discovery

Definition 4 (Causal discovery problem) Let $\mathcal{D}$ be a dataset over variables $\mathbf{V}$, and $(\mathcal{G}, P)$ the CN that generated $\mathcal{D}$. The *causal discovery problem* consists of recovering the CG $\mathcal{G}$ from data $\mathcal{D}$ *and* prior knowledge [22, 52, 59].

Given the true CG $\mathcal{G} = (\mathbf{V}, \mathbf{E})$, a common assumption is that of *causal sufficiency*, although implausible in many scenarios [29]. Causal sufficiency holds when no *hidden*, i.e., unmeasured, variable is a cause of at least two other variables.

There are two main categories of causal discovery algorithms, namely *score-based* and *constraint-based* [22, 57, 59]. Score-based algorithms aim to find a DAG $\mathcal{G}^*$ that maximizes a goodness-of-fit function, called the *score-function*. Constraint-based algorithms exploit conditional independence tests to learn the graph's adjacencies and to orient as many as possible. There is no agreement on which of the two categories is the best in the scientific literature. However, Scutari et al. [49] show that constraint-based approaches perform better in small sample size settings. Another important aspect is that causal discovery algorithms typically do not recover a unique CG, but an MEC, when only observational data are available. Experimental data usually give additional

Table A1 Examples of prior knowledge constraints.

#	Constraint	Input Example	Constraint
1	Required directed	$X \to Y$	$\mathcal{G}$ must contain $X \to Y$.
2	Required undirected	$X - Y$	X and Y must be adjacent in $\mathcal{G}$.
3	Forbidden directed	$X \not\to Y$	$\mathcal{G}$ must not contain $X \to Y$.
4	Forbidden undirected	$X \not- Y$	X and Y must not be adjacent in $\mathcal{G}$.
5	Relaxed partial order	Tier 1: $\{X_1\}$, Tier 2: $\{X_2, X_3\}$	Nodes in tier i are not causes of those in tier j, with $i > j$.
6	Strict partial order	Tier 1: $\{X_1\}$, Tier 2: $\{X_2, X_3\}$	The same as rule 5, but edges within nodes in the same tier are forbidden.
7	Root node	$\not\to X$	X must not have parents in $\mathcal{G}$.
8	Sink node	$X \not\to$	X must not have children in $\mathcal{G}$.

information, allowing us to identify a unique CG [9, 27], but their collection is not always feasible. In this work, we focus on the setting where only observational data are available and resort to expert knowledge for obtaining a unique CG [19].

A.3 Formalizing Expert Knowledge

Expert knowledge, or *prior knowledge*, is any information that constraints or guides the causal discovery algorithm, typically obtainable from domain experts. There are several approaches to formalize expert knowledge (Section B); in this paper, we focus on *hard* constraints [14], namely, rules on the presence/absence of specific edges in the CG. We denote prior knowledge in this form as $\mathcal{K}$ [35]. Some constraints are reported in Table A1; it is straightforward to include each rule in constraint-based CD methods [35]. Note that rules 4-8 can be transformed into rule 3. When the number of nodes increases, the edges grow exponentially, making edge-by-edge elicitation unfeasible. On the contrary, partial orders such as rules 5 and 6 are easier to elicit because they describe partial temporal orderings.

A.4 Estimating Uncertainty in Causal Discovery

Once the causal discovery problem has been solved, the quality of the recovered CG must be assessed. This task may be accomplished by *bagging* [21], which essentially aggregates a set of CGs learned from sampled subsets of the original dataset. Specifically, let $\mathcal{G}$ be the true CG and $\mathcal{H}$ the one learned from data $\mathcal{D}$. *Bootstrap aggregation*, abbreviated in *bagging*, produces many samples $\{\mathcal{D}_1, \ldots, \mathcal{D}_k\}$ from $\mathcal{D}$, with or without repetition of data items. In this way, it simulates datasets with a slightly different joint distribution of variables. The sampling technique can also be conditioned on some variables' values whenever their prevalence is low in $\mathcal{D}$. Then, a CG is learned from each sample, resulting in a set of graphs $\{\mathcal{H}_1, \ldots, \mathcal{H}_k\}$. The probability that X and Y are adjacent in $\mathcal{G}$ is estimated as their adjacency frequency in $\{\mathcal{H}_1, \ldots, \mathcal{H}_k\}$. A threshold t can be chosen so that edges in $\mathcal{H}$ are kept only if their probability is

higher than t. The value of t may be estimated from data, and a common choice is based on the cumulative distribution function of the empirical probabilities [48]. The probability of edge directions is estimated similarly and selected whenever above 0.5. The empirical probabilities behave as a posterior distribution of $\mathcal{G}$ given $\mathcal{D}$. We can then evaluate the robustness of $\mathcal{H}$ and its generalizability to new data.

Appendix B Related Work

Expert knowledge elicitation for causal discovery is based mainly on constraining the graph to be learned, such as setting lists of required and forbidden edges. Several works [11, 12, 35] provide methods to include these rules in causal discovery algorithms. Gonzales et al. [23] also consider a degree of uncertainty in expert knowledge. However, their multi-step pipeline is not iterative, and the final phase only queries experts for edge directions. Brouillard et al. [10] describe a different approach, where variables are assigned a specific category, and edge directions are constrained to consider that category. Borboudakis and Tsamardinos [5] consider rules on a graph's paths, by merging a given causal graph and a set of logical rules while dealing with inconsistency between these two. The assumption of having full access to these *hard* constraints [14] as input to the algorithm is often unrealistic. Constantinou et al. [14] introduce *soft* rules to guide the algorithmic search by bounding the graph space. Authors of [3, 24, 26] consider an expert-given graph *prior*. Borboudakis and Tsamardinos [6] propose a similar methodology, where the prior is based on experts' beliefs of causal and associative relationships. Amirkhani et al. [2] examine the case of both uncertain and heterogeneous knowledge. Regrettably, edges may lose their causal semantics when exploiting soft constraints [23].

The line of *active learning* [28] proposes iterative methods for knowledge elicitation. Kitson and Constantinou [28] modify the *Tabu* algorithm [7] to request new information dynamically. Masegosa and Moral [33] develop an iterative system capable of requesting knowledge as needed. The *cost* of inquiries and the level of experts' reliability are considered. However, experts do not accompany the entire graph construction, and prior knowledge is only requested to establish edges deemed unreliable from the data. Mascaro et al. [32] describe how a causal graph is obtainable by relying uniquely on experts. Sousa et al. [51] develop a process that includes domain experts and tackles the data scarcity problem. Yet, their workflow is not iterative, and prior knowledge must obey the statistical properties of data.

The class of time-varying graphical models [47] gives a valid option to represent longitudinal dynamics—for instance, Nogueira et al. [39] model time series from the healthcare domain. However, longitudinal data are characterized by discrete, and not continuous, sequences of events and right-censoring. Sheidaei et al. [50] leverage the *dynamic Bayesian network* (DBN) framework [37] to model longitudinal data. DBNs assume we observe the system state at points equally spaced in time. Then, due to heterogeneous granularities in patients' observations, *small* time windows should be considered. This may lead to an over-parametrized model that hinders its learning and tractability. Moreover, the DBN *stationarity* is commonly assumed over time, which is not the case under evolving mechanisms. The framework of *continuous time*

Bayesian network (CTBN) [38] allows us to describe a system's evolution without being constrained to discrete time. CTBNs are flexible and promising event-based models. Regrettably, this framework is underdeveloped to obtain a CTBN from longitudinal data [1, 8, 18]. The approaches of [34] and [58] merge causal discovery with longitudinal analysis [36]. They exploit survival analysis to predict the censored time-to-events, then causal discovery is performed on the completed data. Regrettably, they do not consider multiple events and longitudinal dynamics.

Appendix C Longitudinal Data

C.1 Notation

Longitudinal data consist of repeated observations of a cohort of individuals over time, hence are characterized by a set of *events* $\mathcal{E}$ [36]. An *event* $E \in \mathcal{E}$ is any fact that may happen during the patient's observation period. For instance, it can be the onset of the disease, its progression, or its prognosis. Let $t = 0$ be the initial time point of any observation period [54]. The time at which an event E occurs is said to be its *time-to-event* T^E. Longitudinal studies aim to model the event-related *risk* and the cumulative probability distribution of T^E. A subject i is observed up to time T_i, throughout a set of *follow-up* (FUP) visits which may record event occurrences [53]. The step function $E_i(t) = \mathbb{I}(T_i^E \leq t)$ describes whether E occurred before time t for the i-th subject, with $0 \leq t \leq T_i$ and $\mathbb{I}$ the *indicator function*, which is 1 if the argument is *true* and 0 otherwise. Note that the T_i's are commonly heterogeneous across patients; even if an event has not occurred within the last recorded FUP, we cannot conclude that it will not happen afterwards. This is particularly true in the case of *competitive events*. For instance, the death event S is competitive when it prevents the observation of subsequent events unless it is the primary focus of the study. The meaning of T_i^S and $S_i(t)$ follow from T_i^E and $E_i(t)$. A time-to-event T_i^E is said to be *right-censored* if $T_i^E > T_i$. We say the i-th patient is *right-censored* when T_i^S is right-censored; this may happen when the subject leaves the study. Classical analysis for these data involves, for instance, the Kaplan-Meier estimate [44] or the Cox Proportional Hazard model [20] in case patients' covariates and multiple events are considered.

C.2 *DataWindowing*: A Practical Approach to Time-Varying Mechanisms

We present the *DataWindowing* algorithm, a preprocessing step to causal discovery tailored for observational longitudinal data.

Let $\mathcal{D}$ be a dataset in which each row is associated with a subject $i \in \{1, \ldots, N\}$ and each column is associated with the observations of either a non-event variable, i.e., determined at $t = 0$ for any i, or an *event variable*. An *event variable* related to $E \in \mathcal{E}$ is the stochastic process $E(t)$ expressing the probability that $T_E < t$, with $t > 0$. The related column in $\mathcal{D}$ is:

$$E(T_i) = \{E_i(T_i) \mid i \in \{1, \ldots, N\}\} \tag{C4}$$

Algorithm 1 *DataWindowing* algorithm.

Input: dataset $\mathcal{D}$, time τ.
Output: dataset $\mathcal{D}^\tau$.

```
1:  function DataWindowing(D, τ)
2:      D^τ ← copy(D)
3:      for all i ∈ rows(D) do
4:          ε ← ε(i, τ)
5:          if T_i < τ − ε and S_i(T_i) = 0 then
6:              Remove i-th subject from D^τ
7:              Continue
8:          for all E ∈ ℰ do
9:              D^τ[i, E(T_i)] ← E_i(τ)
10:         D^τ[i, S(T_i)] ← S_i(τ)
11:     return D^τ
```

Analogous definitions follow for the stochastic process $S(t)$ and the column $S(T_i)$, related to the death event S. Notice that $E(t)$ depends on t, hence, $E(t)$ is well-defined for the i-th subject only over $t \leq T_i$. Consequently, there may not exist a unique time τ for which $E(\tau)$ is well-defined for *any* subject. However, there are no multiple versions of the same node in a CG. Our proposal tackles the issue by removing the dependency of $E(t)$ and $S(t)$ on t.

Let $\tau > 0$ be a fixed time-point. Algorithm 1 builds a dataset $\mathcal{D}^\tau$ from $\mathcal{D}$, recording whether the events occurred within the time window $[0, \tau]$. Alive subjects observed less than τ are removed from $\mathcal{D}$ in Lines 5-6. The right-censoring justifies this; we cannot know whether they died within $(T_i, \tau]$. On the contrary, all those who died before τ are kept to avoid *selection biases* [25]. Here, $T_i = T_i^S$, so all the event occurrences are recorded at T_i. Each column $E(T_i)$ is substituted with the realisations of the new variable $E(\tau)$ in Line 9. If the i-th patient died within $[0, \tau]$ then $E_i(\tau)$ is set as $E_i(T^S)$; otherwise, $E_i(\tau)$ is set to zero. This way, event variables are given a shared temporal semantics: while $E(t)$ depended on t, now $E(\tau)$ does not. The same argument holds for the death event S and the column $S(T_i)$. We do not distinguish between disease-related death and death from other causes.

Follow-ups may be several months apart. If the last FUP is *near* the selected τ for subject i-th, namely $\tau - \varepsilon \leq T_i \leq \tau$ with ε relatively *small*, then the next FUP will probably be after τ. Hence, including i in $\mathcal{D}^\tau$ is sensible and increases the sample size. Experts may provide a value for ε in Line 4, which depends on (i) τ because FUP frequency usually drops over time, and on (ii) i because high-risk subjects are checked more often than low-risk ones.

The choice of τ requires a trade-off and relies on experts. The smaller the time window is, the higher the related sample size. On the other hand, time is needed for the events to develop: a small τ could hide some cause-effect relationships, because event variables may influence each other from before to after τ. Multiple values $\tau_1 < \ldots < \tau_k$ result in different datasets $\mathcal{D}^{\tau_1}, \ldots, \mathcal{D}^{\tau_k}$ representing the system over overlapping time windows $[0, \tau_1], \ldots, [0, \tau_k]$. Then, the CGs $\mathcal{G}^{\tau_1}, \ldots, \mathcal{G}^{\tau_k}$ can be learned from them through shared or non-shared prior knowledge. When compared, the CG structures

may allow experts to deduce whether, and how, causal mechanisms change over time. Edge removals, additions, and reversals hint that the data-generating process develops over time.

Appendix D Soft Tissue Sarcoma Case Study

We describe the STS data and report the analyses conducted. To do this, we map the steps of Figure 1 to our specific case study. In the following, `Recurrence` indicates the variable (and graph node) associated with local recurrence, while `Metastasis` indicates the variable associated with distant metastasis. The R package *bnlearn* [46] is employed for all the analyses.

Step (1)

The STS dataset belongs to Fondazione IRCCS Istituto Nazionale dei Tumori (INT), Milan, Italy, and has been prospectively collected over the past 30 years. We consider all patients affected by STS who underwent surgery from 2000 to 2020. We exclude all subjects excised before 2000 because the classification of the disease and its treatments have changed. Patients who received an unplanned excision in other institutes and a second, complete excision at INT are included because the re-excision is considered to reset the risks. Hence, $t = 0$ refers to the time of surgery at INT. We select subjects affected by STS of the limbs, arms, and superficial trunk, as causal mechanisms differ in other sites. The final sample includes 2 007 patients.

Steps (2) to (4)

Table D2 lists the selected variables and levels. The reported sample size results from step (7). `Grade` and `Histotype` are proxies of the cancer aggressiveness. Patients who were first excised in other institutes are not differentiated from the others, as surgery at INT *resets* the risk of local recurrence. We exclude the following variables due to their low prevalence in the data and ineffective impact on the CGs globally: `Type of surgery`, `Postoperative complications`, `Metastasis site`, and `Recurrence treatment`. `Age` is discretized in tertiles, while the levels of `Size` are given by expert knowledge based on its role in the disease's natural history. Histotypes reflect previous studies (see Section 3). Treatments, namely `Radiotherapy`, `Chemotherapy`, and *Isolated limb perfusion* (ILP) [15] only refer to whether they were performed or not and are *perioperative*, namely, performed close to surgery. The time of administration, whether before or after surgery, adds no oncological information but scatters data.

Steps (5) to (7)

`Age`, `Size`, `Depth`, and `Margins` are measured at surgery at INT. Specifically, the tumor `Size` is the maximum diameter between the one at first surgery and the one at re-excision. `Recurrence` and `Metastasis`, refer to their first occurrence. We ignore subsequent events because of their low prevalence. Clinicians highlight that imperceptible metastases may also be present at surgery. Thus, temporal sequences of events do not generally suggest any underlying causal relationship. We assume *causal sufficiency* and adopt the discrete Bayesian network framework [30]. Missing values are

Table D2 STS variables and levels resulting from the CD workflow.

Variable and levels	n	%
Total	1 979	-
`Sex`		
Female	880	44.5
Male	1 099	55.5
`Age` (years)		
$x \leq 48$	660	33.4
$48 < x \leq 65$	662	33.5
$x \geq 65$	657	33.2
`Site`		
Upper extremity	281	14.2
Lower extremity	1 257	63.5
Trunk	441	22.3
`Histotype`		
Leiomyosarcoma	186	9.4
DD or pleomorphic liposarcoma	113	5.7
Myxoid liposarcoma	261	13.2
MPNST	112	5.7
Myxofibrosarcoma	333	16.8
Synovial sarcoma	133	6.7
UPS	427	21.6
Vascular sarcoma	54	2.7
Other	360	18.2
`Size` (cm)		
$x \leq 5$	749	37.8
$5 < x \leq 10$	698	35.3
$x > 10$	532	26.9

Variable and levels	n	%
`Depth`		
Superficial	574	29.0
Deep	1 405	71.0
`Grade` (FNCLCC) [55]		
1	387	19.6
2	558	28.2
3	1 034	52.2
`Margins` [45]		
R0	1 719	86.9
R1	258	13.0
R2	2	0.1
`Radiotherapy`		
Done	937	47.3
Not done	1 042	52.7
`Chemotherapy`		
Done	601	30.4
Not done	1 378	69.6
`ILP`		
Done	43	2.2
Not done	1 936	97.8
`Recurrence`		
Yes	188	9.5
No	1 791	90.5
`Metastasis`		
Yes	528	26.7
No	1 451	73.3
`Death`		
Yes	446	22.5
No	1 533	77.5

rare and *completely at random* [47]. Experts fill in some values by logic, and when unfeasible, the data item is excluded from the data. At the end of this step, we are left with 1 979 patients.

Steps (8) to (10)

The meaning of event variables changes depending on $\tau_1, \ldots, \tau_k$ (C). Prior knowledge about concurrent mechanisms helps determine the time windows (more info in the main text). Their values are set to $\tau_1 = 2$, $\tau_2 = 5$, and $\tau_3 = 7$ years. By employing Algorithm 1, we obtain distinct datasets for each time window: $\mathcal{D}^2$ with 1 848 patients, $\mathcal{D}^5$ with 1 579 patients, and $\mathcal{D}^7$ with 1 345 patients. The dataset $\mathcal{D}^2$ includes 294 individuals who developed only `Metastasis`, 52 individuals who developed only `Recurrence`, and 59 individuals who developed both. Similarly, $\mathcal{D}^5$ contains 318 observations of only `Metastasis`, 67 observations of only `Recurrence`, and 75 observations of both. In the case of $\mathcal{D}^7$, we have 314 observations of only `Metastasis`, 63 observations of only `Recurrence`, and 78 observations of both.

Table D3 Final prior knowledge elicitation in the STS case study.

#	Input	Reason
5	Tier 1: {`Age, Sex`}	Patient's covariates.
	Tier 2: {`Histotype, Grade`}	Tumor characteristics.
	Tier 3: {`Site, Size, Depth`}	Physical traits.
	Tier 4: {`Margins, ILP, Chemo., Radio.`}	Surgery quality and therapies.
	Tier 5: {`Recurrence, Metastasis`}	Re-occurrence of disease.
	Tier 6: {`Death`}	Disease prognosis.
3	`Size` $\not\to$ `Site`	Tumor size cannot alter its site.
	`Margins` $\not\to$ `ILP`	`ILP` is performed before surgery.

Steps (11) to (14)

The elicited prior knowledge is shared among all causal discovery tasks, i.e., it constrains $\mathcal{G}^2, \mathcal{G}^5$, and $\mathcal{G}^7$ equally. The final version of prior knowledge is described in Table D3. Due to its stability and explainability, we choose the *PC-Stable* algorithm [13] for causal discovery. As conditional independence test, we exploit the *permutation mutual information* with conditional Monte Carlo simulation (MC-MI) test [16] Its significance threshold is set to $\alpha = 0.05$ and the number of permutations to 5 000, as supported by [56]. MC-MI output differs slightly from run to run because of the stochasticity of permutations. Hence, bagging is required to assess the CGs' stability. Bagging performs 200 resamplings [17], by stratifying on variables `Metastasis` and `Recurrence` due to their low prevalence.

Steps (15) to (24)

Many iterations and refinements are required to obtain three consensus networks (see Section 3). As the study ended, 13 external clinicians assessed the obtained CGs through a questionnaire. The questionnaire focused on the presence/absence of paths related to the effect of therapies and the disease behavior. In conclusion, the causal sufficiency assumption is argued.

D.1 Questionnaire

For reproducibility purposes, we report the questions in the STS questionnaire. The types and fields of answers are reported for each question.

1. *"Email (optional)"*. Free-text input.
2. *"Workplace"*. Multiple choice (Oncology / Surgery / Pathological Anatomy / Radiotherapy / Radiology / Other:___).
3. *"Specialization"*. Free-text input (e.g., "Oncologist", "Surgeon", etc.).
4. *"Years of experience on soft tissue sarcoma"*. Numerical input.
5. *"I took part in this project"*. Multiple choice (Yes / No / Other: ___).
6. *"Positive margins (R1 or R2) can cause local recurrence"*. 5-point Likert scale + optional comment.

7. *"Positive margins (R1 or R2) can cause distant metastasis"*. 5-point Likert scale + optional comment.
8. *"Positive margins (R1 or R2) are also caused by biological variables (e.g., site, histology). If yes, specify which variables"*. 5-point Likert scale + optional comment.
9. *"The decision of whether to administer chemotherapy at diagnosis of a STS of a limb or superficial trunk is based on the observation of"*. Multiple choice (Never / Rarely / Sometimes / Often / Always) for the following factors + optional comment.
   - Histology
   - Grading
   - tumor size
   - tumor depth
   - Patient age
   - Other therapies already performed
   - Postoperative margins
10. *"The decision of whether to administer radiotherapy at diagnosis of a STS of a limb or superficial trunk is based on the observation of"*. Multiple choice (Never / Rarely / Sometimes / Often / Always) for the following factors + optional comment.
    - Histology
    - Grading
    - tumor size
    - tumor depth
    - Patient age
    - Other therapies already performed
    - Postoperative margins
11. *"Chemotherapy prevents local recurrence"*. 5-point Likert scale + optional comment.
12. *"Chemotherapy prevents distant metastasis"*. 5-point Likert scale + optional comment.
13. *"Radiotherapy prevents local recurrence"*. 5-point Likert scale + optional comment.
14. *"Radiotherapy prevents distant metastasis"*. 5-point Likert scale + optional comment.
15. *"Chemotherapy improves patient prognosis in terms of survival"*. 5-point Likert scale + optional comment.
16. *"Radiotherapy improves patient prognosis in terms of survival"*. 5-point Likert scale + optional comment.
17. *"In the short term (i.e., first 2 years), distant metastasis can cause local recurrence"*. 5-point Likert scale + optional comment.
18. *"In the long term (i.e., beyond 5 years), distant metastasis can cause local recurrence"*. 5-point Likert scale + optional comment.
19. *"A local recurrence can cause distant metastasis"*. 5-point Likert scale + optional comment.

20. *"In the short term, local recurrence can cause patient death"*. 5-point Likert scale + optional comment.
21. *"In the long term, local recurrence can cause patient death"*. 5-point Likert scale + optional comment.
22. *"In the short term, certain histologies and grading cause distant metastasis"*. 5-point Likert scale + optional comment.
23. *"In the short term, certain histologies and grading cause local recurrence"*. 5-point Likert scale + optional comment.
24. *"In the long term, histology and grading do not cause local recurrence"*. 5-point Likert scale + optional comment.

References

[1] Alessandro B (2024) Structure learning and knowledge extraction with continuous time bayesian network. PhD thesis, URL https://hdl.handle.net/20.500.14242/161745

[2] Amirkhani H, Rahmati M, Lucas PJF, et al (2017) Exploiting experts' knowledge for structure learning of bayesian networks. IEEE Trans Pattern Anal Mach Intell 39(11):2154–2170. https://doi.org/10.1109/tpami.2016.2636828

[3] Angelopoulos N, Cussens J (2008) Bayesian learning of bayesian networks with informative priors. Ann Math Artif Intel 54(1–3):53–98. https://doi.org/10.1007/s10472-009-9133-x

[4] Bareinboim E, Correa JD, Ibeling D, et al (2022) On Pearl's Hierarchy and the Foundations of Causal Inference, ACM, pp 507–556. https://doi.org/10.1145/3501714.3501743

[5] Borboudakis G, Tsamardinos I (2012) Incorporating causal prior knowledge as path-constraints in bayesian networks and maximal ancestral graphs. In: Proceedings of the 29$^{\text{th}}$ International Coference on International Conference on Machine Learning. Omnipress, Madison, WI, USA, ICML'12, pp 427–434

[6] Borboudakis G, Tsamardinos I (2013) Scoring and searching over bayesian networks with causal and associative priors. In: Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence. AUAI Press, Arlington, Virginia, USA, UAI'13, pp 102–111

[7] Bouckaert RR (1995) Bayesian belief networks: from construction to inference. PhD thesis

[8] Bregoli A, Scutari M, Stella F (2021) A constraint-based algorithm for the structural learning of continuous-time bayesian networks. Int J Approx Reason 138:105–122. https://doi.org/https://doi.org/10.1016/j.ijar.2021.08.005, URL https://www.sciencedirect.com/science/article/pii/S0888613X21001304

[9] Brouillard P, Lachapelle S, Lacoste A, et al (2020) Differentiable causal discovery from interventional data. In: Larochelle H, Ranzato M, Hadsell R, et al (eds) Advances in Neural Information Processing Systems, vol 33. Curran Associates, Inc., pp 21865–21877, URL https://proceedings.neurips.cc/paper_files/paper/2020/file/f8b7aa3a0d349d9562b424160ad18612-Paper.pdf

[10] Brouillard P, Taslakian P, Lacoste A, et al (2022) Typing assumptions improve identification in causal discovery. In: Schölkopf B, Uhler C, Zhang K (eds) Proceedings of the First Conference on Causal Learning and Reasoning, Proceedings of Machine Learning Research, vol 177. arXiv, pp 162–177, URL https://proceedings.mlr.press/v177/brouillard22a.html

[11] de Campos CP, Zeng Z, Ji Q (2009) Structure learning of bayesian networks using constraints. In: Proceedings of the 26th Annual International Conference on Machine Learning. ACM, New York, NY, USA, ICML '09, p 113–120, https://doi.org/10.1145/1553374.1553389, URL https://doi.org/10.1145/1553374.1553389

[12] de Campos LM, Castellano JG (2007) Bayesian network learning algorithms using structural restrictions. Int J Approx Reason 45(2):233–254. https://doi.org/10.1016/j.ijar.2006.06.009

[13] Colombo D, Maathuis MH (2014) Order-independent constraint-based causal structure learning. J Mach Learn Res 15(1):3741–3782

[14] Constantinou AC, Guo Z, Kitson NK (2023) The impact of prior knowledge on causal structure learning. Knowl Inf Syst 65(8):3385–3434. https://doi.org/10.1007/s10115-023-01858-x

[15] Creech O, Krementz ET, Ryan RF, et al (1958) Chemotherapy of cancer: Regional perfusion utilizing an extracorporeal circuit. Ann Surg 148(4):616–632. https://doi.org/10.1097/00000658-195810000-00009

[16] Edwards D (2000) Introduction to Graphical Modelling. Springer New York, https://doi.org/10.1007/978-1-4612-0493-0

[17] Efron B, Tibshirani R (1998) An introduction to the bootstrap, [nachdr.] edn. No. 57 in Monographs on statistics and applied probability, Chapman & Hall, Boca Raton, Fla. [u.a.], originally publ. by Chapman & Hall

[18] Engelmann N, Linzner D, Koeppl H (2020) Continuous time Bayesian networks with clocks. In: III HD, Singh A (eds) Proceedings of the 37th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 119. PMLR, pp 2912–2921, URL https://proceedings.mlr.press/v119/engelmann20a.html

[19] Fenton N, Neil M (2018) Risk Assessment and Decision Analysis with Bayesian Networks. Chapman and Hall/CRC, https://doi.org/10.1201/b21982

[20] Fox J (2008) An R-and S-Plus companion to applied regression, [nachdr.] edn. Sage, Thousand Oaks, Calif. [u.a.]

[21] Friedman N, Goldszmidt M, Wyner A (1999) Data analysis with bayesian networks: a bootstrap approach. In: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence. arXiv, San Francisco, CA, USA, UAI'99, pp 196–205

[22] Glymour C, Zhang K, Spirtes P (2019) Review of causal discovery methods based on graphical models. Front Genet 10. https://doi.org/10.3389/fgene.2019.00524

[23] Gonzales C, Journe A, Mabrouk A (2022) A hybrid algorithm for learning causal networks using uncertain experts' knowledge. In: Salmerón A, Rumı´ R (eds) Proceedings of The 11th International Conference on Probabilistic Graphical Models, Proceedings of Machine Learning Research, vol 186. PMLR, pp 241–252, URL https://proceedings.mlr.press/v186/gonzales22a.html

[24] Heckerman D, Geiger D, Chickering DM (1995) Learning bayesian networks: The combination of knowledge and statistical data. Mach Learn 20(3):197–243. https://doi.org/10.1007/bf00994016

[25] Hernán MA, Sterne JAC, Higgins JPT, et al (2024) A structural description of biases that generate immortal time. Epidemiology 36(1):107–114. https://doi.org/10.1097/ede.0000000000001808

[26] Imoto S, Higuchi T, Goto T, et al (2004) Combining microarrays and biological knowledge for estimating gene networks via bayesian networks. J Bioinform Comput Biol 02(01):77–98. https://doi.org/10.1142/s021972000400048x

[27] Jaber A, Kocaoglu M, Shanmugam K, et al (2020) Causal discovery from soft interventions with unknown targets: Characterization and learning. In: Larochelle H, Ranzato M, Hadsell R, et al (eds) Advances in Neural Information Processing Systems, vol 33. Curran Associates, Inc., pp 9551–9561, URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6cd9313ed34ef58bad3fdd504355e72c-Paper.pdf

[28] Kitson NK, Constantinou AC (2023) Causal discovery using dynamically requested knowledge. https://doi.org/10.2139/ssrn.4620804

[29] Kocaoglu M, Shanmugam K, Bareinboim E (2017) Experimental design for learning causal graphs with latent variables. In: Guyon I, Luxburg UV, Bengio S, et al (eds) Advances in Neural Information Processing Systems, vol 30. Curran Associates, Inc., URL https://proceedings.neurips.cc/paper_files/paper/2017/file/291d43c696d8c3704cdbe0a72ade5f6c-Paper.pdf

[30] Koller D, Friedman N (2010) Probabilistic graphical models, [nachdr.] edn. Adaptive computation and machine learning, MIT Press, Cambridge, Mass. [u.a.],

includes bibliographical references and index

[31] Lauritzen SL (1996) Graphical Models. Oxford University PressOxford, https://doi.org/10.1093/oso/9780198522195.001.0001

[32] Mascaro S, Wu Y, Woodberry O, et al (2023) Modeling covid-19 disease processes by remote elicitation of causal bayesian networks from medical experts. BMC Med Res Methodol 23(1). https://doi.org/10.1186/s12874-023-01856-1

[33] Masegosa AR, Moral S (2013) An interactive approach for bayesian network learning using domain/expert knowledge. Int J Approx Reason 54(8):1168–1181. https://doi.org/10.1016/j.ijar.2013.03.009

[34] Mbogu HM, Nicholson CD (2024) Data-driven root cause analysis via causal discovery using time-to-event data. Comput Ind Eng 190:109974. https://doi.org/10.1016/j.cie.2024.109974

[35] Meek C (1995) Causal inference and causal explanation with background knowledge. In: Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, UAI'95, pp 403–410

[36] Miller RG (2011) Survival analysis, wiley classics library ed (online-ausg.) edn. Wiley classics library, Wiley-Interscience, New York, includes bibliographical references and index. - Electronic reproduction; Palo Alto, Calif; ebrary; 2011; Available via World Wide Web; Access may be limited to ebrary affiliated libraries

[37] Murphy KP (2002) Dynamic bayesian networks: representation, inference and learning. PhD thesis, Computer Science, URL https://ibug.doc.ic.ac.uk/media/uploads/documents/courses/DBN-PhDthesis-LongTutorail-Murphy.pdf

[38] Nodelman U, Shelton CR, Koller D (2002) Continuous time bayesian networks. In: Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, UAI'02, p 378–387

[39] Nogueira AR, Abreu Ferreira C, Gama J (2022) Temporal Nodes Causal Discovery for in Intensive Care Unit Survival Analysis, Springer International Publishing, pp 587–598. https://doi.org/10.1007/978-3-031-16474-3_48

[40] Pearl J (1995) From Bayesian Networks to Causal Networks, Springer US, pp 157–182. https://doi.org/10.1007/978-1-4899-1424-8_9

[41] Pearl J (2020) The book of why, first trade paperback edition edn. Basic Books, New York, literaturverzeichnis: Seite 377-404

[42] Pearl J (2021) Causal inference in statistics, reprinted with revisions edn. Wiley, Chichester, literaturverzeichnis: Seite 127-131

[43] Peters J (2017) Elements of causal inference. Adaptive computation and machine learning, The MIT Press, Cambridge, Massachusetts

[44] Rich JT, Neely JG, Paniello RC, et al (2010) A practical guide to understanding kaplan-meier curves. Otolaryngology–Head and Neck Surgery 143(3):331–336. https://doi.org/10.1016/j.otohns.2010.05.007

[45] Sambri A, Caldari E, Fiore M, et al (2021) Margin assessment in soft tissue sarcomas: Review of the literature. Cancers 13(7):1687. https://doi.org/10.3390/cancers13071687

[46] Scutari M (2010) Learning bayesian networks with the bnlearn r package. J Stat Softw 35(3). https://doi.org/10.18637/jss.v035.i03

[47] Scutari M (2020) Bayesian network models for incomplete and dynamic data. Stat Neerl 74(3):397–419. https://doi.org/10.1111/stan.12197

[48] Scutari M, Nagarajan R (2013) Identifying significant edges in graphical models of molecular networks. Artif Intell Med 57(3):207–217. https://doi.org/10.1016/j.artmed.2012.12.006

[49] Scutari M, Graafland CE, Gutiérrez JM (2019) Who learns better bayesian network structures: Accuracy and speed of structure learning algorithms. Int J Approx Reason 115:235–253. https://doi.org/10.1016/j.ijar.2019.10.003

[50] Sheidaei A, Foroushani AR, Gohari K, et al (2022) A novel dynamic bayesian network approach for data mining and survival data analysis. BMC Med Inform Decis Mak 22(1). https://doi.org/10.1186/s12911-022-02000-7

[51] Sousa HS, Prieto-Castrillo F, Matos JC, et al (2018) Combination of expert decision and learned based bayesian networks for multi-scale mechanical analysis of timber elements. Expert Syst Appl 93:156–168. https://doi.org/10.1016/j.eswa.2017.09.060

[52] Spirtes P, Glymour C, Scheines R (2001) Causation, Prediction, and Search. The MIT Press, https://doi.org/10.7551/mitpress/1754.001.0001

[53] Suchmacher M, Geller M (2012) Practical biostatistics. Elsevier, Amsterdam [u.a.], https://doi.org/10.1016/c2011-0-04190-x, includes bibliographical references and index

[54] Therneau TM, Grambsch PM (2000) Multiple Events per Subject, Springer New York, pp 169–229. https://doi.org/10.1007/978-1-4757-3294-8_8

[55] Trojani M, Contesso G, Coindre JM, et al (1984) Soft-tissue sarcomas of adults; study of pathological prognostic variables and definition of a histopathological grading system. Int J Cancer 33(1):37–42. https://doi.org/10.1002/ijc.2910330108

[56] Tsamardinos I, Borboudakis G (2010) Permutation Testing Improves Bayesian Network Learning, Springer Berlin Heidelberg, pp 322–337. https://doi.org/10.1007/978-3-642-15939-8_21

[57] Vowels MJ, Camgoz NC, Bowden R (2022) D'ya like dags? a survey on structure learning and causal discovery. ACM Computing Surveys 55(4):1–36. https://doi.org/10.1145/3527154

[58] Štajduhar I, Dalbelo-Bašić B (2010) Learning bayesian networks from survival data using weighting censored instances. J Biomed Inform 43(4):613–622. https://doi.org/10.1016/j.jbi.2010.03.005

[59] Zanga A, Ozkirimli E, Stella F (2022) A survey on causal discovery: Theory and practice. Int J Approx Reason 151:101–129. https://doi.org/10.1016/j.ijar.2022.09.004