# Causal Model Discovery in Cancer Guided by Cellular Pathways[⋆]

Rodrigo Henrique Ramos[1,2,3][0000−0002−0786−5387],
Adenilso Simao[1][0000−0002−1454−2607], and
Mohammad Reza Mousavi[2][0000−0002−4869−6794]

[1] University of São Paulo, Av. Trab. São Carlense, 400, São Carlos, São Paulo, Brazil
[2] King's College London, Strand, London WC2R 2LS, England
[3] Federal Institute of São Paulo, Estrada Municipal Paulo Eduardo de Almeida, São Carlos - São Paulo, Brasil
ramos@ifsp.edu.br

**Abstract.** Cancer is a genetic disease caused by mutations accumulated over time that corrupt cell pathways. Factors such as smoking, sunlight exposure, and age can influence the emergence of the disease and contribute to its heterogeneity. Identifying sub-groups within cancer populations is crucial for tailored treatments, given varied genetic responses to therapies. The increasing availability of (public) cancer data enables the development of many statistical studies exploring the relationships between mutations, pathways, and risk factors. Causal discovery can enhance these relationships into causal associations. Our paper presents an approach for investigating causal models in cancer, considering patient data, tumor stage, and pathways while addressing mutational diversity through patient sub-groups. Utilizing data from lung cancer, we use cancer driver genes and pathways to group patients associated with 12 cellular functions. We design and apply a causal discovery method to extract causal associations from the patient level to the genomic level, offering an overview of the entire set of patients. We also construct causal models for the 12 sub-groups of patients, exploring the different causal associations found within the same cancer type. Our results considering all patients show expected causal associations between patients and tumor data and new insights for causality among cellular pathways, especially Programmed Cell Death and Metabolism of Proteins. Also, sub-groups of patients within the same cancer type have distinct causal associations. These findings highlight the importance of considering mutational heterogeneity and personalized characteristics when exploring the complex landscape of causal discovery for cancer.

**Keywords:** Causality · Causal Discovery · Cancer Genomics · Cancer Patient Data · Cellular Pathways.

[⋆] This research was supported by King's College London, University of São Paulo, São Paulo Research Foundation (FAPESP), and the Brazilian Federal Foundation for Support and Evaluation of Graduate Education (CAPES). M.R. Mousavi was partially supported by UKRI and EPSRC (grants EP/V026801/2 and EP/Y005244/1.

## 1 Introduction

The causes for the emergence and spread of cancer are a central question in cancer research [31]. Causality [26], as a formal approach to identifying relations beyond simple correlation, has evolved over the last few decades. This has led to the creation of methods and algorithms to establish causal links between random variables in order to create causal models (CMs) [14]. Causal discovery from observational data goes beyond statistical correlation alone, e.g., by excluding confounding factors that influence both the purported causes and effects. In the worst case, this may require combinatorial calculations that are exponential in the number of variables [21, 33]. The computational cost of causal discovery often limits the number of random variables analysed [14]. This limitation is a problem in the context of cancer research, where it is necessary to consider information from large patient-, tumor-, pathway- and gene datasets.

The temporal cellular processes of genetic mutations, driver mutations, pathway corruption, and cancer development form a causal network [16, 31]. However, there are also indirect factors for cancer development, such as phenotype, environment, lifestyle, and diet [3, 8, 35]. When analysing the direct and indirect factors for cancer development, it is also important to consider sub-groups of patients to address the mutational heterogeneity of subcohorts in the same cancer type [4, 20]. Examining the multifaceted aspects of causality in cancer can improve our understanding of the disease's initiation, progression, and spread [31]. To the best of our knowledge, no compositional causal analysis technique has been employed to explore those direct and indirect links. By "compositional", we refer to a (potentially mechanizable) technique that can divide the dataset into sub-groups, discover causal relations in an efficient manner, and then generalise the results up to the global data set as much as warranted by a causal analysis, while annotating causal relations that are specific to sub-groups.

In this study, we develop an approach for discovering causal models in cancer. We consider patient data, tumor stage, and pathways while addressing mutational heterogeneity through patient sub-groups. For each patient, we only use mutations in known cancer driver genes. We group the patients considering biological pathways that intersect with mutated drivers, enabling meaningful association between patients and cellular processes. This reduction from mutations in all genes to only driver genes and from drivers to pathways significantly decreases the number of variables, enabling the execution of the causal model discovery algorithm while creating sub-groups of the whole population. We create causal models for the sub-groups of patients in order to explore the nuances and specificities of causal associations found within the same cancer type. Since the same patients are associated with more than one group, we also create a causal model considering all patients and their associations with the pathways.

Analysis of the entire patient cohort revealed expected causal connections between patients and tumor data, while also uncovering novel insights into causal relationships within cellular pathways, particularly with Programmed Cell Death and Metabolism of Proteins. Since we made the causal discovery from observational data without prior knowledge, the expected associations found increased the method's credibility. Additionally, causal models constructed for patient sub-groups unveiled distinct and recurrent causal associations. A consensus causal model was then developed to consolidate findings from the sub-groups, emphasizing recurring patterns. These findings underscore the significance of accounting for mutational diversity and personalized characteristics while exploring the intricate landscape of cancer causality. By creating causal models that incorporate patient, tumor, and genomic data from both the entire cohort and sub-groups, this study offers complementary insights into varying levels of causal associations within cancer.

To summarise, the contributions of this paper are twofold:

1. Designing a scalable and compositional causal discovery pipeline for a composition of hetergenous dataset with three major components: variable reduction, decomposition, and aggregation; and
2. Applying the methodology to a composition of three types of cancer datasets and obtaining novel causal models and derive novel insights from them.

The remainder of this paper is organized as follows: Section 2 provides the background information for the rest of the paper. Section 3 present our methodology for a scalable and compositional causal discovery pipeline. Section 4 presents the results of applying our methodology and presents the gained insights. Section 5 concludes the paper.

## 2 Background

### 2.1 Cancer data

Cancer is a complex disease that emerges from the intricate interplay of mutations that disrupt the normal regulation of genes and cellular pathways [16]. Advances in DNA sequencing methods have led to the availability of a large volume of cancer data [24, 25]. Databases and portals such as TCGA and cBioPortal provide public data from multiple cancer studies. The specific information may change from study to study and also depends on the type of cancer (i.e. smoking in lung cancer and menopause in breast cancer) [27]. At a global level, all studies tend to present at least three types of datasets:

1. datasets with de-identified and anonymized *patient information*,
2. datasets with *tumor information* for each patient, and
3. datasets containing aggregated *mutation information*, including detailed genomic analysis of mutations.

The National Institutes of Health (NIH) offers detailed information about format of the last type of datasets, called the Mutation Annotation Format (MAF) [18]. A MAF file lists all mutations found in each gene for each patient [23]. Most of these mutations are unrelated to cancer development, thus called passenger mutations [22]. On the other hand, driver mutations are responsible for cancer development and progression [16, 22]. The search for drivers led to the creation of curated lists of known driver genes, such as NCG [10], and the development of many computational methods that generally use the MAF file combined with protein networks and pathways [6, 7, 9, 17, 30].

### 2.2 Pathways

Cellular pathways are complex networks of molecular reactions that govern essential biological processes within cells. These pathways encompass a wide array of functions, such as signal transduction, DNA replication, metabolism, protein synthesis, and others. Researchers rely on databases such as Reactome, KEGG, and WikiPathways to study these pathways. Understanding these pathways is crucial for deciphering cellular behaviour's complexities and elucidating the mechanisms underlying diseases [19, 36].

Reactome is a curated knowledgebase of biological pathways that focuses on human biology and offers a hierarchical organization of cellular processes, enabling researchers to explore interconnected molecular events at different levels of complexity [13]. Reactome uses the concept of Super Pathways to organize its pathways hierarchically. There are 27 Super Pathways (SPs) representing the main cellular process and 2,654 sub-pathways related to the main processes. SPs have different sets of genes and display complex and distinct topology when modelled as networks [28]. Figure 1 shows the Reactome pathway hierarchy for the SP Programmed Cell Death. Programmed Cell Death is a main cellular process and has 215 genes. This process is divided into two sub-pathways, Apoptosis with 181 genes and Regulated Necrosis with 61, meaning these two pathways share 27 genes. The SP hierarchical tree spreads in depth and width. The Programmed Cell Death has 41 sub-pathways, with smaller sets of genes and more specific cellular functions at the bottom.

### 2.3 Causal discovery

Causal discovery is a crucial pursuit in a variety of fields. Its goal is to reveal the causal relationships that underlie the observed data [26]. This involves using methods and algorithms to infer causal structures from observational or experimental data, often relying on probabilistic reasoning and statistical dependencies [14]. By identifying causal relationships, researchers can uncover the mechanisms that drive phenomena of interest, leading to informed decision-making and predictive modelling [26]. Ultimately, causal discovery enriches our comprehension of complex systems and has numerous practical applications.

Identifying causal relationships from observational data is complex and computationally intensive, especially for large-scale or high-dimensional data [14]. One of the most prominent approaches in this field is the Peter-Clark (PC) algorithm, which identifies directed acyclic graphs (DAGs) by testing conditional independence assumptions iteratively [21]. The algorithm initiates with a complete undirected graph, representing the maximum potential set of edges without prior knowledge of the underlying causal relationships. The algorithm then refines the graph by increasing conditional independence tests with edge removal

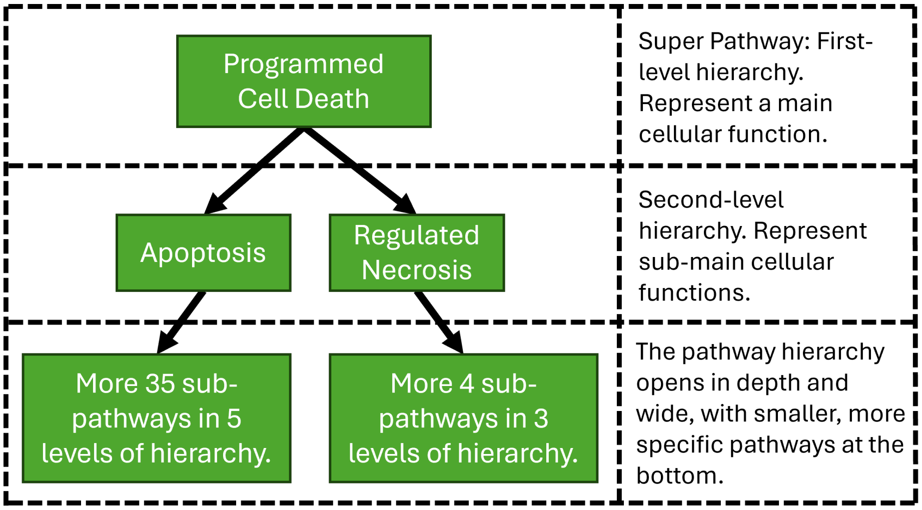

Programmed Cell Death
Apoptosis
Regulated Necrosis
More 35 sub-pathways in 5 levels of hierarchy.
More 4 sub-pathways in 3 levels of hierarchy.
Super Pathway: First-level hierarchy. Represent a main cellular function.
Second-level hierarchy. Represent sub-main cellular functions.
The pathway hierarchy opens in depth and wide, with smaller, more specific pathways at the bottom.

**Fig. 1.** Reactome Super Pathways Hierarchy. Reactome organizes its pathways in a tree-like structure. With the Super Pathway representing a main cellular function as root and more generic and smaller pathways as descendants.

based on the results. This process ultimately leads to a minimal graph, called skeleton, that reflects the observed conditional independence relationships in the data. After obtaining the skeleton, the PC algorithm orients edges by applying rules based on the skeleton topology to determine causality between variables. A detailed explanation of the rules is presented in the work of Le Thuc Duy [21]. The algorithm transforms the skeleton into a directed acyclic graph (DAG) representing causal relationships among variables. When an edge cannot be oriented, it remains undirected, indicating correlations between variables that may be affected by common causes or confounding factors but not a direct causal relationship. When edges remain undirected, the resulting graph may not necessarily be acyclic. As the number of variables increases, the complexity of conditional independence tests increases exponentially. Scalability has been traditionally improved by parallelization and approximation strategies, but it still faces challenges related to its time complexity in high-dimensional data [14, 21].

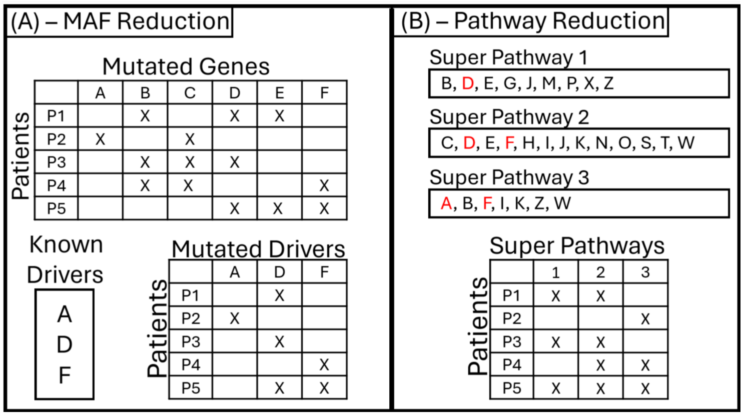
## 3 Causal Discovery Pipeline

Our pipeline starts by exploring and preprocessing the patient and tumor data. We removed outliers and discretised continuous variables as usual. Due to space limitation, we included the details of these steps in the Appendix. After the preprocessing, we divide our data preparation into three steps: variable reduction, decomposition, and aggregation. Finally we apply the causal discovery algorithm on the reduced, decomposed, and aggregated outcomes and compose them back again.

### 3.1 Variable Reduction and Decomposition

These two steps, respectively, address the following two requirements: 1) We must reduce the number of variables to run the causal discovery algorithm. 2) We must address the mutation heterogeneity by creating sub-groups of patients.

Considering variables of interest in cancer study from lung cancer [5], in the patient dataset, there are five variables: Age, Sex, Smoking History, Smoking Packing Years (SPY), and Status (alive or deceased). The tumor dataset has the Stage variable. Given the time complexity of causal discovery methods, these six variables are manageable if we want to analyse just patient and tumor data. The problem arises when we consider the hundreds of cellular pathways and the approximately 20,000 protein-coding genes [32]. The MAF file from the study mentioned above has 17,959 protein-coding genes. Since most of those mutations are passengers, the first step is to consider only known drivers. The NCG database presents a list of 3,347 drivers. This procedure removes all false positive genes for cancer (passengers) at the expense of removing possible unknown drivers. Figure 2-(A) exemplifies the passenger removal from the MAF file, a step that reduces the number of genes from 17,959 to 3,347. This reduction is significant and also removes false positives, but it is still unmanageable for combinatorial causal discovery methods. We assume a causal relation in cancer between drivers genes and SPs [16], then we associate each patient with the Reactome's SPs by intersecting their genes with patients' mutated drivers.



| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| P1 | | X | | X | X | |
| P2 | X | | X | | | |
| P3 | | X | X | X | | |
| P4 | | X | X | | | X |
| P5 | | | | X | X | X |

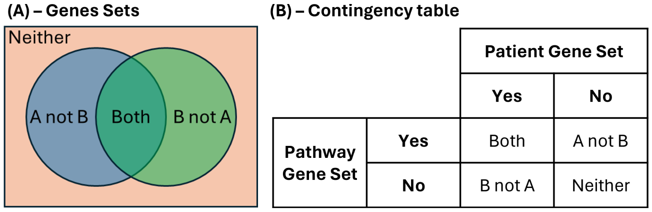
| | A | D | F |
|---|---|---|---|
| P1 | | X | |
| P2 | X | | |
| P3 | | X | |
| P4 | | | X |
| P5 | | X | X |

| | 1 | 2 | 3 |
|---|---|---|---|
| P1 | X | X | |
| P2 | | | X |
| P3 | X | X | |
| P4 | | X | X |
| P5 | X | X | X |

**Fig. 2.** Reducing the number of variables. In (A), we use a curated list of known cancer driver genes from NCG to remove passenger genes while keeping all the patients. In (B), we use the Mutated Drivers table from (A) to decompose the data and associate each patient to pathways harbouring one or more altered drivers.

### 3.2 Aggregation

Figure 2-(B) shows how we intersect the Mutated Drivers table from Figure 2-(A) with Reactome's SPs. P1 in Figure 2-(A) has mutations on the genes B, D, and E. The pathways in Figure 2-(B) intersect with these three genes, but since only D is a known driver, we associated P1 with the two first pathways. We need to test the statistical significance of these intersections between pathway genes and patients drivers. For each patient and for each SP, we test if the intersection between genes is statistically significant (p-value $\leq 0.05$). If it is, we add the patient in the SP group. If the SP group contains less than 221 statistically significant patients (20% of all patients), we search in the SP hierarchy for a sub-pathway that statistically significantly comprises more than 220 patients. We use the Fisher Exact Test [1] to establish the statistical significance between gene sets, in a similar way used to find significant associations among two sets of genes in mutual exclusivity or co-occurrence [2, 12]. We calculate $(Neither * Both)/(AnotB * BnotA)$, where $A$ is the patient genes, and $B$ is the SP or sub-pathway genes. To calculate *Neither*, we use all genes in the MAF file. Figure 3-(A) illustrates the intersection and Figure 3-(B) the contingency table used in the statistical test.



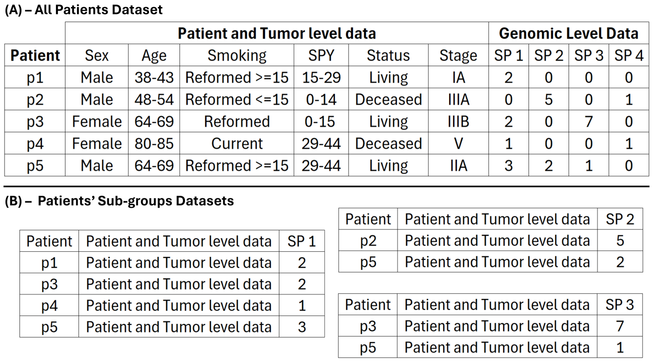
| | | Patient Gene Set | |
|---|---|---|---|
| | | Yes | No |
| Pathway Gene Set | Yes | Both | A not B |
| | No | B not A | Neither |

**Fig. 3.** Significant intersection. $A$ is the patient set of genes, and $B$ is the SP or sub-pathway set genes. *Neither* is the set of all other genes found in the MAF.

After this first step, we considerably reduce the number of variables in the genomic level from 17,959 to, at most, 27. This reduction makes the application of combinatorial causal discovery algorithms possible. The association of patients and SPs or sub-pathways using a statistical test enables the meaningful creation of sub-groups of patients. These groups address the mutational heterogeneity found in cancer and have biological significance.

### 3.3 Datasets for Causal Model Discovery

In this third step, we prepare the datasets to discover causal models. We expand the final table from Figure 2 regarding genomic level information to include patient and tumor level data. This dataset stores information about all patients and SPs or sub-pathways. Since the patients may be grouped in multiple pathways, we can extract causal links between them together with patient and tumor data, offering an overview of the cancer type. It is possible to divide the dataset considering the pathways, creating sub-groups of patients sharing driver genes in the same biological process. These smaller datasets keep the six patient and tumor level data and one pathway at a time. Figure 4 exemplify the datasets.



**(A) – All Patients Dataset**

| | Patient and Tumor level data | | | | | | Genomic Level Data | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Patient** | Sex | Age | Smoking | SPY | Status | Stage | SP 1 | SP 2 | SP 3 | SP 4 |
| p1 | Male | 38-43 | Reformed >=15 | 15-29 | Living | IA | 2 | 0 | 0 | 0 |
| p2 | Male | 48-54 | Reformed <=15 | 0-14 | Deceased | IIIA | 0 | 5 | 0 | 1 |
| p3 | Female | 64-69 | Reformed | 0-15 | Living | IIIB | 2 | 0 | 7 | 0 |
| p4 | Female | 80-85 | Current | 29-44 | Deceased | V | 1 | 0 | 0 | 1 |
| p5 | Male | 64-69 | Reformed >=15 | 29-44 | Living | IIA | 3 | 2 | 1 | 0 |

**(B) – Patients' Sub-groups Datasets**

| Patient | Patient and Tumor level data | SP 1 |
|---|---|---|
| p1 | Patient and Tumor level data | 2 |
| p3 | Patient and Tumor level data | 2 |
| p4 | Patient and Tumor level data | 1 |
| p5 | Patient and Tumor level data | 3 |

| Patient | Patient and Tumor level data | SP 2 |
|---|---|---|
| p2 | Patient and Tumor level data | 5 |
| p5 | Patient and Tumor level data | 2 |

| Patient | Patient and Tumor level data | SP 3 |
|---|---|---|
| p3 | Patient and Tumor level data | 7 |
| p5 | Patient and Tumor level data | 1 |

**Fig. 4.** Datasets for causal model discovery. In (B), we suppress the variable's names and the sub-group for SP 4 to save space in the plot, but all datasets have six variables: Sex, Age, Smoking, SPY, Status, and Stage.

Figure 4-(A) shows the dataset for all patients and SPs or sub-pathways that significantly group patients. The variables can be divided into two types: Patient and Tumor Level Data, and Genomic Level Data. The latter shows how many drivers each patient has on each SP or sub-pathway. For the SP or sub-pathway be included as column, it must significantly group 221 patients (20% of all patients) or more. For this reason, it is not guaranteed that all 27 SPs or sub-pathways will be in the dataset.

Figure 4-(B) displays datasets for subgroups of patients with driver mutations occurring in the same biological process. These sub-groups provide a means of categorizing the MAF file based on mutational diversity. It is possible to investigate the causal associations of each group and develop a consensus causal model, which shows all the associations identified in each group, placing more weight on recurring edges. Figure 5 illustrates these models.

In Figure 5 (B), and (C), we exemplify how different types of patient and tumor data may interact. Dashed edges indicate a strong correlation that persists through independent sets but not causation. Direct edges indicate causation. In Figure 5 (C) the edge's weight helps identify the strength of correlation or causation in different sub-groups of patients.

## 4 Results

Using data from lung cancer, we found 5 SP and 7 sub-pathways that significantly grouped more than 220 patients (20% of the 1,107 patients after preprocessing).

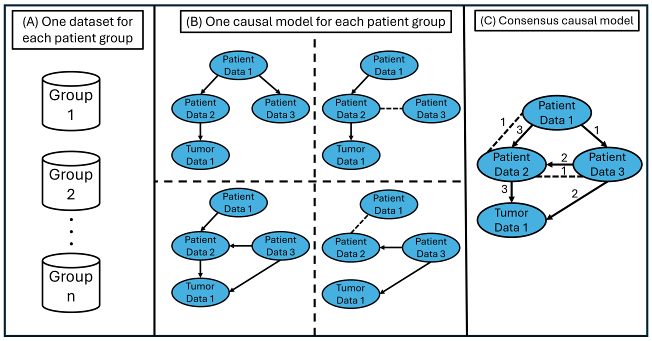

(A) One dataset for each patient group
Group 1
Group 2
Group n
(B) One causal model for each patient group
Patient Data 1
Patient Data 2
Patient Data 3
Tumor Data 1
(C) Consensus causal model
1
3
1
2
1
3
2

**Fig. 5.** Sub-groups and consensus causal model. In (A), we show the patients datasets from Figure 3. In (B) we create one causal model for each patient group that may vary on the edges. In (C), we combine all causal models, adding weights for repeating edges.

Table 1 shows the pathway names and number of patients in each group. Sub-pathways are indicated by "- -", with the SP name before the "- -" and the sub-pathway name after it. Each group in Table 1 harbours, on average, 35% of all patients. (For details, we have included Figure 14 in the Appendix, which shows the intersection of patients in one group (rows) with all other groups (columns).) Cell-Cell Communication has 339 patients, of which 29% are also on Cellular responses to stimuli - - Oncogene Induced Senescence (with 277 patients), and 64% are in Development Biology (with 589 patients).

**Table 1.** Patient groups. All SP or sub-pathways, indicated by "- -", which significantly grouped more than 220 patients.

| **Super Pathway groups** | **Patients** |
|---|---|
| Cell-Cell communication | 339 |
| Cellular responses to stimuli - - Oncogene Induced Senescence | 277 |
| Chromatin organization - - PKMTs methylate histone lysines | 237 |
| Developmental Biology | 589 |
| Extracellular matrix organization | 518 |
| Gene expression (Transcription) - - Transcriptional regulation by RUNX3 | 246 |
| Immune System - - Dectin-2 family | 279 |
| Metabolism of proteins - - SUMOylation | 330 |
| Neuronal System | 428 |
| Programmed Cell Death - - Activation of NOXA and translocation | 656 |
| Signal Transduction | 518 |
| Vesicle-mediated transport - - Scavenging by Class H Receptors | 235 |

### 4.1 Sub-groups and Consensus Causal Models

With 12 groups of patients associated with cellular functions, we executed the PC algorithm from the TETRAD library [29] and plotted the networks using the NetworkX library [15]. Section *"Causal Diagrams"* on the Appendix presents the 12 causal models. Figure 6 shows 4 causal models with distinct topology. Each causal model is from a sub-group of patients and has a dataset equal to Figure 4-(B), with 6 variables from patients and tumor, and 1 variable representing the SP or sub-pathway.

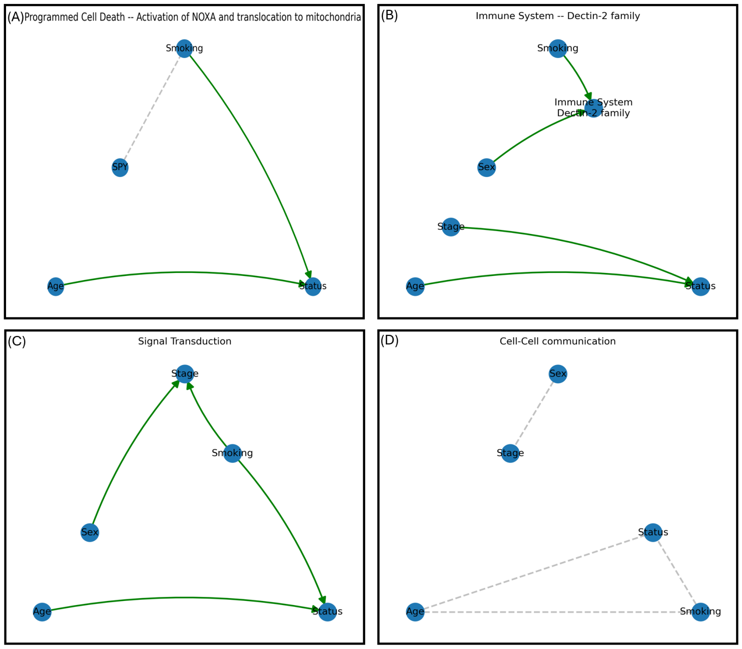


**Fig. 6.** Four causal models. Grey dashed edges represent significant correlation, and solid green edges causation. The sub-plots (C) and (D) shows the causal models for two SP, Signal Transduction and Cell-Cell Communication. The sub-plots (A) and (B) shows causal models for sub-pathways.

It can be observed that not all 7 variables are included in the causal models, and the nodes have few edges. This is because the PC algorithm searches for conditional independent sets, so the absence of nodes and edges is significant. Age and Smoking are linked only in the causal model for Cell-Cell Communication, but in the other three, this connection was eliminated because some combination of the other 5 variables made them independent. Age and Status are associated in all models, meaning this association persist independently of any combination of the other variables.

In Figure 6-(A), the group of patients with driver mutations on a sub-pathway of Programmed Cell Death have a causal association between Smoking and Status, and Age and Status. In these patients, age and the habit of smoking cause death or survival. In Figure 6-(D), these three variables are also associated, but not in a causal way. In Figure 6-(B), the corresponding node to the sub-pathway Dectin-2 Family within the hierarchy of the SP Immune System is caused by Smoking and Sex. The V-Structure $Smoking \rightarrow Pathway \leftarrow Sex$ indicates that the corruption of the pathways by mutated drivers depends at the same time of smoking and sex. This finding brings insights into the complex dependencies in cancer development and the interplay of direct and indirect factors in cancer. The rest of this section further explores these associations.

All 12 causal models are unique, indicating that the process of grouping patients to reduce complexity and address mutational heterogeneity successfully extracted distinct sub-groups of patients. Albeit unique, the causal models have re-occurring edges. Figure 7 shows a consensus causal model with all associations found in the 12 models, with edge weight indicating the occurrences numbers.

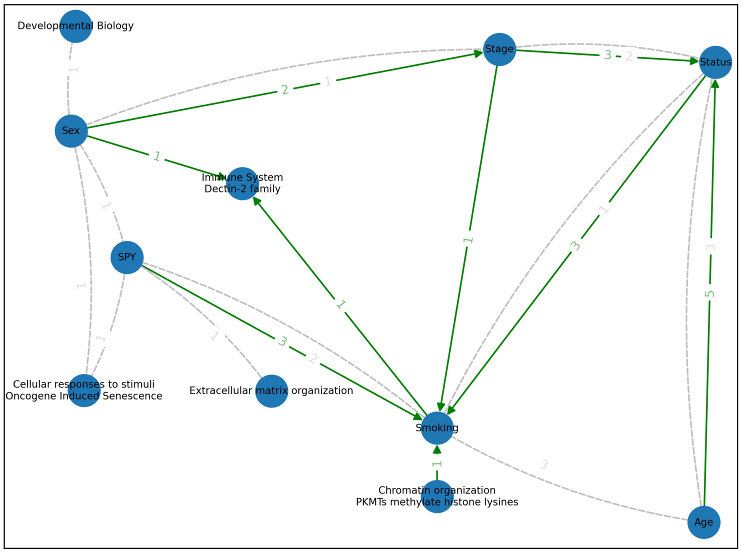

Developmental Biology
Stage
Status
Sex
Immune System
Dectin-2 family
SPY
Cellular responses to stimuli
Oncogene Induced Senescence
Extracellular matrix organization
Smoking
Chromatin organization
PKMTs methylate histone lysines
Age
1
2
1
3
2
1
1
1
1
1
1
3
1
3
5
1
3
2
1
3

**Fig. 7.** Consensus Causal Model. Grey dashed edges represent significant correlation, and solid green edges are causation. Edge weight shows the edge occurrences in the 12 sub-groups causal models.

In Figure 7, Age and Status appear 8 times, 3 as correlation and 5 as Age causing Status. We understand that patient age impacts the chance of survival, and this association was found in 8 causal models. We also see another expected association when Stage cause Status in 3 cases and is correlated in 2. We can also question some causal edge orientations. Status is correlated with Smoking in 1 case, and in 3 cases, Status causes Smoking, also Stage causes Smoking in 1 case. In both cases, we believe the orientation should be inverted. The PC algorithm extracts causal associations from observational data without previous knowledge and expects the distributions to be Gaussian. These few inaccurate edge orientations can be manually fixed by cancer specialists to improve the model's accuracy. The consensus causal model in Figure 7 offers an aggregated summary for causal associations found in sub-groups of patients, and place Smoking and Status as key variables among the others due the number of connections.

### 4.2 All Patient's Causal Model

The lung cancer study used in this paper has 1,107 patients after our preprocessing. Each patient appears on average in 4.5 groups, with a median of 4 and a standard deviation of 2. Since each group represents a cellular process, we can explore the causal association among pathways. Figure 8 presents two causal models created using all patients, considering the dataset in Figure 4-(A). Figure 8-(A) was constructed using only data from the patient and tumor level. Figure 8-(B) adds data from the genomic level. Comparing the patient and tumor variables in the two models, the only difference is the absence of the correlation between Sex and Stage in Figure 8-(B). Some combination of pathways made them conditional independent. On the other hand, any combination of pathways made the other patients and tumor variables independent. It is important to point out the relevance of Smoking in connecting the genomic level to the patient and tumor level.

Age and Smoking are associated in 3 causal models for sub-groups of patients and are not associated with SPY in any of them. In the causal model for all patients, Age and Smoking are intermediated by SPY, meaning that once you have information about the SPY, one does not need the Age to know about the time a patient smokes (i.e. the Smoking variable). Status is conditionally independent from all other nodes once we have information about the patient's Age and tumor Stage. Overall, variables from the patient and tumor level show the expected causal associations with each other, which increases the method's credibility when analysing the association at the genomic level.

Figure 8-(B) shows us the corruption of the pathway within the SP Programmed Cell Death by mutated drivers genes is caused by the time the patient smokes and also by the corruption of 3 other cellular processes. Corruption in a sub-pathway within the Metabolism of Protein is also caused by 3 pathways but is not directly or indirectly caused by Cellular responses to stimuli – Oncogene Induced Senescence in Figure 8-(B). The interplay of associations and disassociations with pathways may be related to the known cancer phenomena of co-occurrence and mutual exclusivity among driver genes. Both phenomena have applications in cancer therapy [11]. Figure 8 presents an intricate network of causal associations from the data at different levels, giving us insights into the complex landscape of cancer causality.

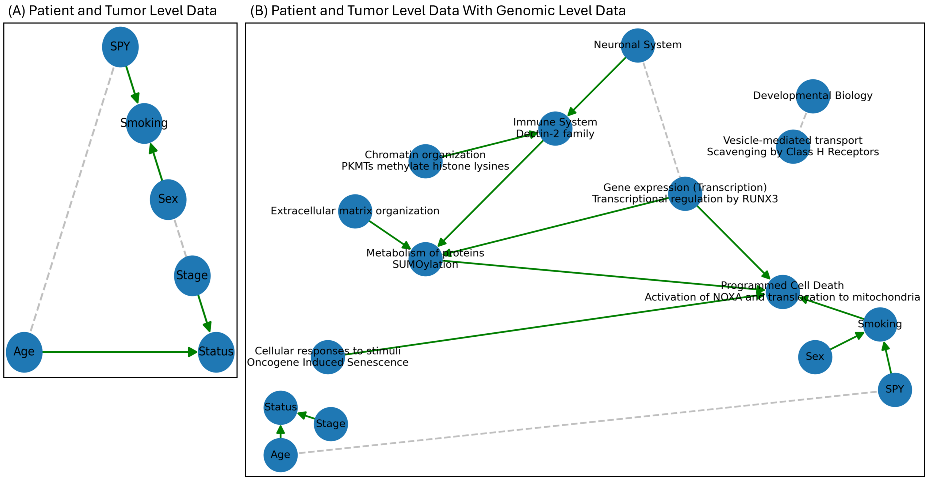

(A) Patient and Tumor Level Data
SPY
Smoking
Sex
Stage
Age
Status
(B) Patient and Tumor Level Data With Genomic Level Data
Neuronal System
Developmental Biology
Immune System
Dectin-2 family
Vesicle-mediated transport
Scavenging by Class H Receptors
Chromatin organization
PKMTs methylate histone lysines
Gene expression (Transcription)
Transcriptional regulation by RUNX3
Extracellular matrix organization
Metabolism of proteins
SUMOylation
Programmed Cell Death
Activation of NOXA and translocation to mitochondria
Smoking
Sex
Cellular responses to stimuli
Oncogene Induced Senescence
SPY
Status
Stage
Age

**Fig. 8.** All patient causal models. (A) shows a causal model with only patients and tumor level data. We see expected causal association among variables. (B) shows a causal model considering also the genomic-level data. We observe causal associations among pathways, which brings insights into the complex interactions of pathways in cancer.

### 4.3 Discussion

The methodology developed in this study enabled the creation of a causal discovery pipeline in cancer. The variable reduction, decomposition, and aggregation steps offer a scalable approach to deal with causal discovery on multidimensional datasets from complementary levels (patients, tumor, and genomic). The use of cancer drivers genes and pathways reduced the number of variables from 17,959 to 12, removed false positives and increased the biological significance of sub-groups.

Causal model analysis reinforced known results while unveiling new insights. Expected causal associations between patient and tumor data were observed, thereby validating the methodology's efficacy. Notably, variables such as smoking and status emerged as pivotal nodes to validate causal associations. Novel insights include distinct causal associations within sub-groups of patients, as in the group Immune System – Dectin-2 family, where the sex variable causes the pathway's corruption. However, this association is not found in any other sub-group. Patterns like this highlight the importance of considering mutational heterogeneity for personalized treatments. Also, the causal model for all patients and pathways brings insights into pathways' complex association and disassociation. Programmed Cell Death – Activation of NOXA and translocation to mitochondria is caused by four other variables but is not causing another variable. This pathway's corruption is the final result of a series of causes.

## 5 Conclusion

Our work aimed to create causal models in cancer, considering data from patient, tumor, and genomic levels while addressing mutational heterogeneity with sub-groups of patients. Using data from more than 1,000 lung cancer patients, we integrate biological knowledge, namely cancer driver genes and pathways, to guide the grouping of patients into biologically relevant cohorts. This procedure reduced the number of variables from the genomic level from 17,959 to 12, making possible the execution of a combinatorial causal discovery method. We created 12 causal models from distinct sub-sets of patients. The models have unique and recurring causal associations. We also created a composed causal model considering all 1,107 patients, with 6 variables from patient and tumor level added with 12 cellular pathways.

Causal models for sub-groups of patients highlight the presence of heterogeneity in causal associations within patients from the same cancer type. The construction of a combined causal model summarized the sub-groups associations, emphasizing recurring edges. Notably, variables such as smoking and survival emerge as critical nodes, exhibiting a higher number of causal associations with other variables in the model.

The causal model with all patients and variables provides an overview of the cancer type. Variables at the patient and tumor levels showed expected causal associations, increasing the method's credibility. Associations and disassociation among pathways might be related to the phenomenon of mutual exclusivity and co-occurrence of cancer driver genes.

In addition to an approach to join cancer data from different levels, the paper bring insights into the complex dependencies in cancer development and the interplay of direct and indirect factors in cancer. In the future, we plan to apply this approach to other types of cancer, comparing causal models from distinct and similar types of cancer. Providing a generic methodology that allows for reusing data across neighbouring cohorts and similar disease types remains an interesting avenue for research. Furthermore, we intend to use the causal models we have developed to perform further causal inference analysis, deepening our understanding of the underlying mechanisms of cancer.

**Acknowledgments.** We acknowledge the informative and constructive comments of anonymous reviewers. M.R. Mousavi has been partially supported by by the UKRI Trustworthy Autonomous Systems Node in Verifiability, grant reference EP/V026801/2, the EPSRC project on Verified Simulation for Large Quantum Systems (VSL-Q), grant reference EP/Y005244/1 and the EPSRC projecton Robust and Reliable Quantum Computing (RoaRQ), Investigation 009, grant reference EP/W032635/1. Also King's Quantum grants provided by King's College London are gratefully acknowledged.

**Disclosure of Interests** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Agresti, A.: Categorical data analysis, vol. 792. John Wiley & Sons (2012)
2. Ahmed, R., Erten, C., Houdjedj, A., Kazan, H., Yalcin, C.: A network-centric framework for the evaluation of mutual exclusivity tests on cancer drivers. Frontiers in genetics **12**, 746495 (2021)
3. Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter, P.: The preventable causes of cancer. In: Molecular Biology of the Cell. 4th edition. Garland Science (2002)
4. Amatya, A.K., Fiero, M.H., Bloomquist, E.W., Sinha, A.K., Lemery, S.J., Singh, H., Ibrahim, A., Donoghue, M., Fashoyin-Aje, L.A., de Claro, R.A., et al.: Subgroup analyses in oncology trials: regulatory considerations and case examples. Clinical Cancer Research **27**(21), 5753–5756 (2021)
5. Campbell, J.D., Alexandrov, A., Kim, J., Wala, J., Berger, A.H., Pedamallu, C.S., Shukla, S.A., Guo, G., Brooks, A.N., Murray, B.A., et al.: Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. Nature genetics **48**(6), 607–616 (2016)
6. Cutigi, J.F., Evangelista, A.F., Reis, R.M., Simao, A.: A computational approach for the discovery of significant cancer genes by weighted mutation and asymmetric spreading strength in networks. Scientific reports **11**(1), 1–10 (2021)
7. Cutigi, J.F., Evangelista, A.F., Simao, A.: Approaches for the identification of driver mutations in cancer: A tutorial from a computational perspective. Journal of Bioinformatics and Computational Biology **18**(03), 2050016 (2020)
8. Danaei, G., Vander Hoorn, S., Lopez, A.D., Murray, C.J., Ezzati, M.: Causes of cancer in the world: comparative risk assessment of nine behavioural and environmental risk factors. The Lancet **366**(9499), 1784–1793 (2005)
9. Dimitrakopoulos, C.M., Beerenwinkel, N.: Computational approaches for the identification of cancer genes and pathways. Wiley Interdisciplinary Reviews: Systems Biology and Medicine **9**(1), e1364 (2017)
10. Dressler, L., Bortolomeazzi, M., Keddar, M.R., Misetic, H., Sartini, G., Acha-Sagredo, A., Montorsi, L., Wijewardhane, N., Repana, D., Nulsen, J., et al.: Comparative assessment of genes driving cancer and somatic evolution in non-cancer tissues: an update of the network of cancer genes (ncg) resource. Genome biology **23**(1), 35 (2022)
11. El Tekle, G., Bernasocchi, T., Unni, A.M., Bertoni, F., Rossi, D., Rubin, M.A., Theurillat, J.P.: Co-occurrence and mutual exclusivity: what cross-cancer mutation patterns can tell us. Trends in cancer **7**(9), 823–836 (2021)
12. Gao, J., Aksoy, B.A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S.O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., et al.: Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal. Science signaling **6**(269), pl1–pl1 (2013)
13. Gillespie, M., Jassal, B., Stephan, R., Milacic, M., Rothfels, K., Senff-Ribeiro, A., Griss, J., Sevilla, C., Matthews, L., Gong, C., et al.: The reactome pathway knowledgebase 2022. Nucleic acids research **50**(D1), D687–D692 (2022)
14. Glymour, C., Zhang, K., Spirtes, P.: Review of causal discovery methods based on graphical models. Frontiers in genetics **10**, 524 (2019)
15. Hagberg, A., Swart, P., S Chult, D.: Exploring network structure, dynamics, and function using networkx. Tech. rep., Los Alamos National Lab.(LANL), Los Alamos, NM (United States) (2008)

16. Hanahan, D.: Hallmarks of cancer: new dimensions. Cancer discovery **12**(1), 31–46 (2022)
17. Hristov, B.H., Chazelle, B., Singh, M.: ukin combines new and prior information with guided network propagation to accurately identify disease genes. Cell systems **10**(6), 470–479 (2020)
18. Institute, N.C.: Gdc maf format v.1.0.0, https://docs.gdc.cancer.gov/Data/File_Formats/MAF_Format/
19. Khatri, P., Sirota, M., Butte, A.J.: Ten years of pathway analysis: current approaches and outstanding challenges. PLoS computational biology **8**(2), e1002375 (2012)
20. Kuipers, J., Thurnherr, T., Moffa, G., Suter, P., Behr, J., Goosen, R., Christofori, G., Beerenwinkel, N.: Mutational interactions define novel cancer subgroups. Nature communications **9**(1), 4353 (2018)
21. Le, T.D., Hoang, T., Li, J., Liu, L., Liu, H., Hu, S.: A fast pc algorithm for high dimensional causal discovery with multi-core pcs. IEEE/ACM transactions on computational biology and bioinformatics **16**(5), 1483–1495 (2016)
22. Martínez-Jiménez, F., Muiños, F., Sentís, I., Deu-Pons, J., Reyes-Salazar, I., Arnedo-Pac, C., Mularoni, L., Pich, O., Bonet, J., Kranas, H., et al.: A compendium of mutational cancer driver genes. Nature Reviews Cancer **20**(10), 555–572 (2020)
23. Mayakonda, A., Lin, D.C., Assenov, Y., Plass, C., Koeffler, H.P.: Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genome research **28**(11), 1747–1756 (2018)
24. Meldrum, C., Doyle, M.A., Tothill, R.W.: Next-generation sequencing for cancer diagnostics: a practical perspective. The Clinical Biochemist Reviews **32**(4), 177 (2011)
25. Nair, S.V., Madhulaxmi, Thomas, G., Ankathil, R.: Next-generation sequencing in cancer. Journal of maxillofacial and oral surgery **20**, 340–344 (2021)
26. Pearl, J., et al.: Causality: Models, reasoning and inference. Cambridge, UK: CambridgeUniversityPress **19**(2), 3 (2000)
27. Ramos, R., Cutigi, J., Ferreira, C., Evangelista, A., Simao, A.: Analyzing different cancer mutation data sets from breast invasive carcinoma (brca), lung adenocarcinoma (luad), and prostate adenocarcinoma (prad). In: Anais do XX Simpósio Brasileiro de Computação Aplicada à Saúde. pp. 37–48. SBC (2020)
28. Ramos, R.H., Cutigi, J.F., Oliveira Lage Ferreira, C.d., Simao, A.: Topological characterization of cancer driver genes using reactome super pathways networks. In: Brazilian Symposium on Bioinformatics. pp. 26–37. Springer (2021)
29. Ramsey, J., Andrews, B.: Py-tetrad and rpy-tetrad: A new python interface with r support for tetrad causal search. In: Causal Analysis Workshop Series. pp. 40–51. PMLR (2023)
30. Reyna, M.A., Leiserson, M.D., Raphael, B.J.: Hierarchical hotnet: identifying hierarchies of altered subnetworks. Bioinformatics **34**(17), i972–i980 (2018)
31. Rondeau, E., Larmonier, N., Pradeu, T., Bikfalvi, A.: Characterizing causality in cancer. Elife **8**, e53755 (2019)
32. Salzberg, S.L.: Open questions: How many genes do we have? BMC biology **16**(1), 94 (2018)
33. Squires, C., Uhler, C.: Causal structure learning: A combinatorial perspective. Foundations of Computational Mathematics **23**(5), 1781–1815 (2023)
34. Tamborero, D., Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Kandoth, C., Reimand, J., Lawrence, M.S., Getz, G., Bader, G.D., Ding, L., et al.: Comprehensive identification of mutational cancer driver genes across 12 tumor types. Scientific reports **3**(1), 2650 (2013)

35. Wu, S., Zhu, W., Thompson, P., Hannun, Y.A.: Evaluating intrinsic and non-intrinsic cancer risk factors. Nature communications **9**(1), 3490 (2018)
36. Zhao, K., Rhee, S.Y.: Interpreting omics data with pathway enrichment analysis. Trends in Genetics **39**(4), 308–319 (2023)

## Appendix

### Preprocessing of Patient and Tumor Data

This section presents the preprocessing performed on patient and tumor variables. We show distribution plots for each variable before and after preprocessing and a textual consideration.

Figures for continuous or discrete variables shows acronyms in the title: Q1, Q1, and Q3 are the quartiles, Avg is the average value, and Max and Min are the maximum and minimum values found on the distribution.

**AGE** – In Figure 9, the first figure shows the age distribution of patients. Since most other variables are categorical, we discretize the age into ten groups, as shown in the second figure.

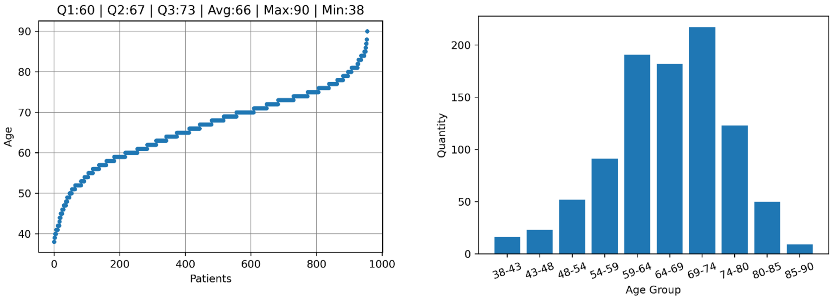


**Fig. 9.** Age original distribution and discretized

**SEX** – No preprocessing was made on the sex variable

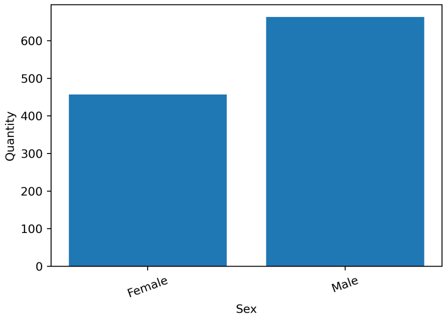


**Fig. 10.** Patient's sex distribution

**SMOKING PACK YEARS (SPY)** – Each number in this variable represents the smoking of one pack of cigarettes a day for one year. For instance, a

SPY of 20 indicates a person smoked one pack per day for 20 years, two packs per day for 10 years, or half a pack per day for 40 years. The two figures on the left show the discrete and categorical distribution before the removal of outliers, and the two figures on the right show the distribution after the removal of outliers. To identify outliers, we remove patients whose values are greater than a threshold, calculated by $Q3 + 4.5 \times IQR$, where Q3 and IQR are the third quartile and the interquartile range of the distribution, a common strategy to remove outliers in cancer datasets [27, 34]. This procedure removed only 6 patients, while keeping the quartiles the same and reducing the maximum value from 240 to 147

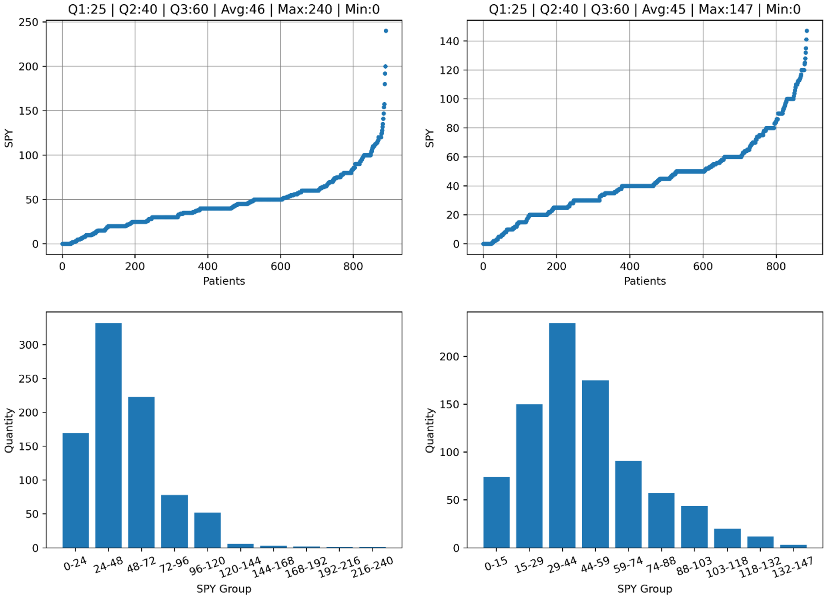


**Fig. 11.** Patient's smoking pack years

**STATUS** – Shows if the patients died or survive during the study that collected the data. No preprocessing made.

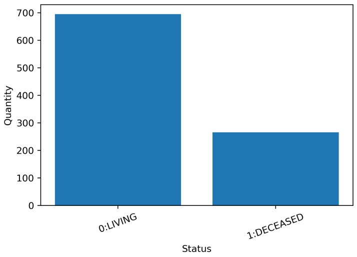


**Fig. 12.** Patient Status

**STAGE** – Represents the cancer stage. The left figure shows the original distribution. Pre-processing consisted of removing stages I, II and III, as they correspond to very small groups.

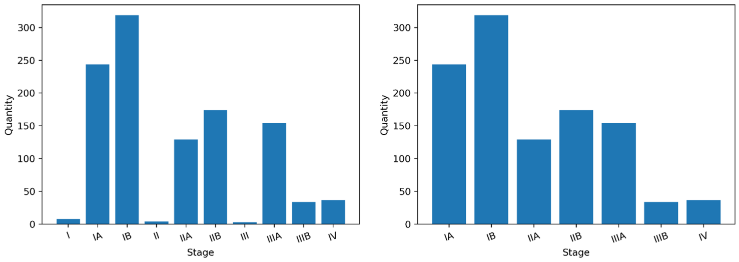
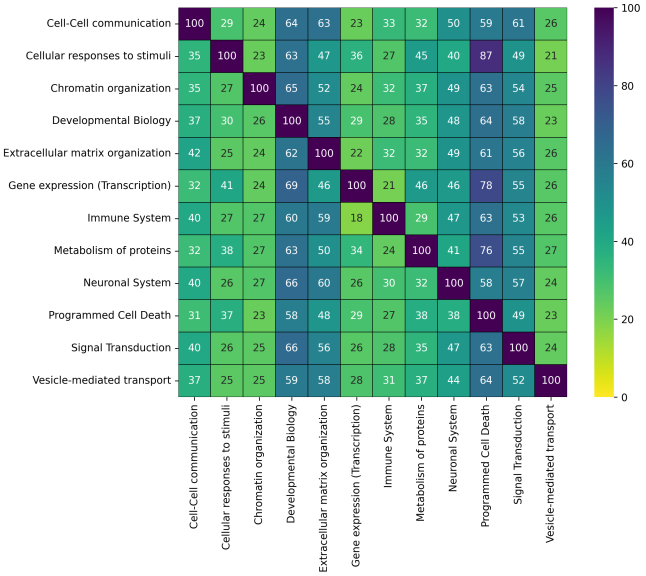


**Fig. 13.** Cancer Stage

## Patient groups' intersections

This section presents the Figure 14 showing how the sub-groups of patients intersect.



| | Cell-Cell communication | Cellular responses to stimuli | Chromatin organization | Developmental Biology | Extracellular matrix organization | Gene expression (Transcription) | Immune System | Metabolism of proteins | Neuronal System | Programmed Cell Death | Signal Transduction | Vesicle-mediated transport |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Cell-Cell communication | 100 | 29 | 24 | 64 | 63 | 23 | 33 | 32 | 50 | 59 | 61 | 26 |
| Cellular responses to stimuli | 35 | 100 | 23 | 63 | 47 | 36 | 27 | 45 | 40 | 87 | 49 | 21 |
| Chromatin organization | 35 | 27 | 100 | 65 | 52 | 24 | 32 | 37 | 49 | 63 | 54 | 25 |
| Developmental Biology | 37 | 30 | 26 | 100 | 55 | 29 | 28 | 35 | 48 | 64 | 58 | 23 |
| Extracellular matrix organization | 42 | 25 | 24 | 62 | 100 | 22 | 32 | 32 | 49 | 61 | 56 | 26 |
| Gene expression (Transcription) | 32 | 41 | 24 | 69 | 46 | 100 | 21 | 46 | 46 | 78 | 55 | 26 |
| Immune System | 40 | 27 | 27 | 60 | 59 | 18 | 100 | 29 | 47 | 63 | 53 | 26 |
| Metabolism of proteins | 32 | 38 | 27 | 63 | 50 | 34 | 24 | 100 | 41 | 76 | 55 | 27 |
| Neuronal System | 40 | 26 | 27 | 66 | 60 | 26 | 30 | 32 | 100 | 58 | 57 | 24 |
| Programmed Cell Death | 31 | 37 | 23 | 58 | 48 | 29 | 27 | 38 | 38 | 100 | 49 | 23 |
| Signal Transduction | 40 | 26 | 25 | 66 | 56 | 26 | 28 | 35 | 47 | 63 | 100 | 24 |
| Vesicle-mediated transport | 37 | 25 | 25 | 59 | 58 | 28 | 31 | 37 | 44 | 64 | 52 | 100 |

**Fig. 14.** Patient groups' intersections. Shows in percentuals how many patients from one group (rows) are also present in other groups. To save space on the plot, we present only the SP name.

## Causal Diagrams

This section presents the causal diagrams for the 12 groups of patients.

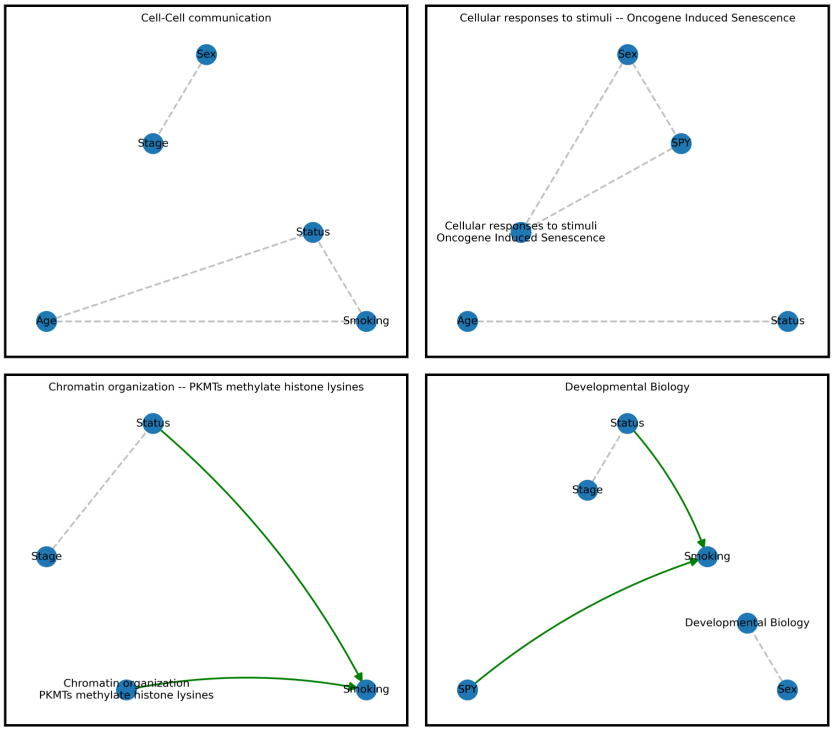

Cell-Cell communication
Sex
Stage
Status
Age
Smoking
Cellular responses to stimuli -- Oncogene Induced Senescence
Sex
SPY
Cellular responses to stimuli
Oncogene Induced Senescence
Age
Status
Chromatin organization -- PKMTs methylate histone lysines
Status
Stage
Chromatin organization
PKMTs methylate histone lysines
Smoking
Developmental Biology
Status
Stage
Smoking
Developmental Biology
SPY
Sex

**Fig. 15.** Causal Models for Cell-Cell communication, Cellular responses to stimuli – Oncogene Induced Senescence, Chromatin organization – PKMTs methylate histone lysines, and Developmental Biology.

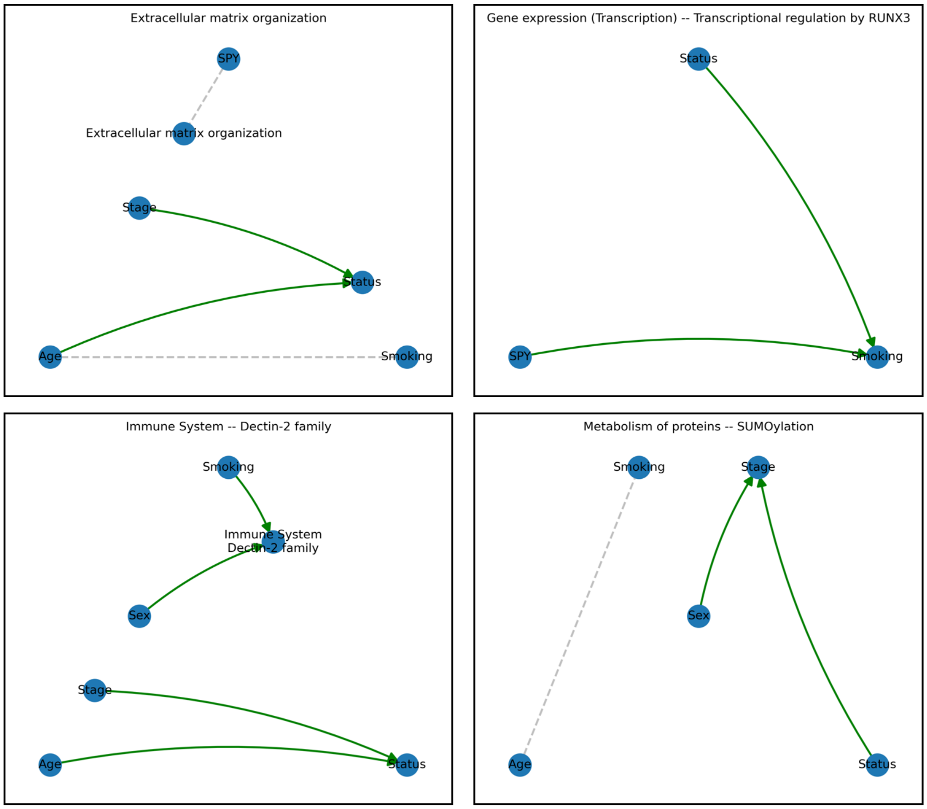

Extracellular matrix organization
SPY
Extracellular matrix organization
Stage
Status
Age
Smoking
Gene expression (Transcription) -- Transcriptional regulation by RUNX3
Status
SPY
Smoking
Immune System -- Dectin-2 family
Smoking
Immune System
Dectin-2 family
Sex
Stage
Age
Status
Metabolism of proteins -- SUMOylation
Smoking
Stage
Sex
Age
Status

**Fig. 16.** Causal Models for Extracellular matrix organization, Gene expression (Transcription) – Transcriptional regulation by RUNX3, Immune System – Dectin-2 family, and Metabolism of proteins – SUMOylation.

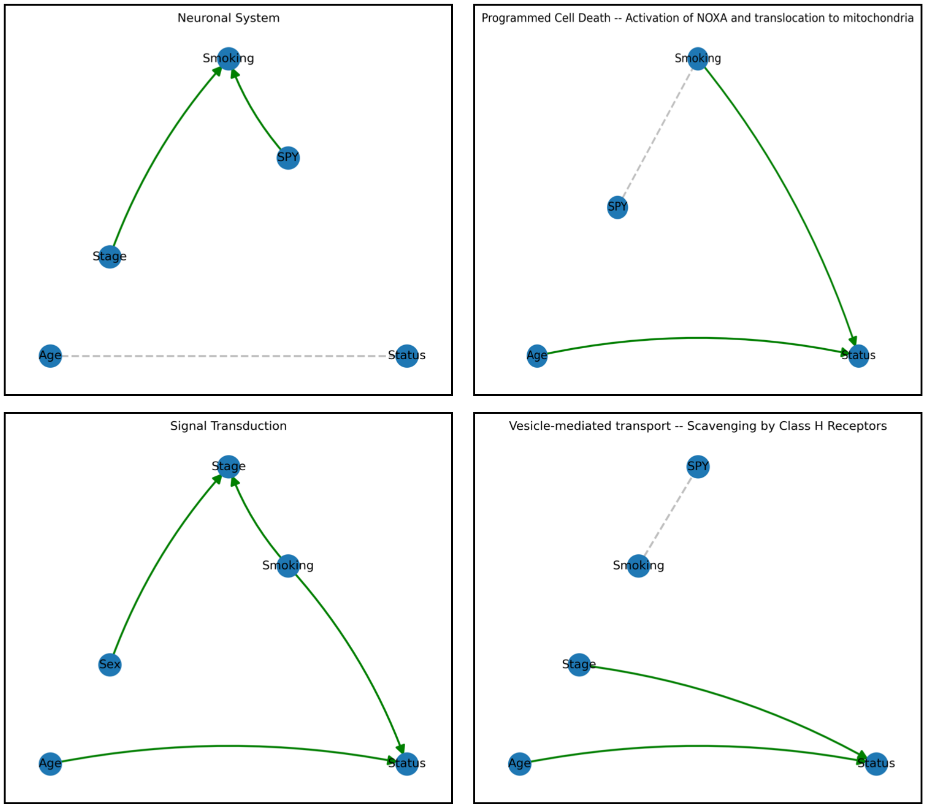

Neuronal System
Smoking
SPY
Stage
Age
Status
Programmed Cell Death -- Activation of NOXA and translocation to mitochondria
Smoking
SPY
Age
Status
Signal Transduction
Stage
Smoking
Sex
Age
Status
Vesicle-mediated transport -- Scavenging by Class H Receptors
SPY
Smoking
Stage
Age
Status

**Fig. 17.** Causal Models for Neuronal System, Programmed Cell Death – Activation of NOXA and translocation to mitochondria, Signal Transduction, and Vesicle-mediated transport – Scavenging by Class H Receptors.