Applying Large Language Models for Causal Structure Learning in Non Small Cell Lung Cancer

Narmada Naik
datma Health Science
Beaverton, OR, USA
narmada@datma.com

Ayush Khandelwal
datma Health Science
Beaverton, OR, USA
ayush@datma.com

Mohit Joshi
datma Health Science
Beaverton, OR, USA
mohit@datma.com

Madhusudan Atre
datma Health Science
Beaverton, OR, USA
atre@datma.com

Hollis Wright
datma Health Science
Beaverton, OR, USA
wrighth@datma.com

Kavya Kannan
datma Health Science
Beaverton, OR, USA
kavyak@datma.com

Scott Hill
datma Health Science
Beaverton, OR, USA
scott@datma.com

Giridhar Mamidipudi
datma Health Science
Beaverton, OR, USA
giri@datma.com

Ganapati Srinivasa
datma Health Science
Beaverton, OR, USA
gans@datma.com

Carlo Bifulco
Earle A. Chiles Research Institute
Providence Cancer Institute
Portland, OR, USA
carlo.bifulco@providence.org

Brian Piening
Earle A. Chiles Research Institute
Providence Cancer Institute
Portland, OR, USA
brian.piening@providence.org

Kevin Matlock
datma Health Science
Beaverton, OR, USA
kevin.matlock@datma.com

***Abstract*—Causal discovery is becoming a key part in medical AI research. These methods can enhance healthcare by identifying causal links between biomarkers, demographics, treatments and outcomes. They can aid medical professionals in choosing more impactful treatments and strategies. In parallel, Large Language Models (LLMs) have shown great potential in identifying patterns and generating insights from text data. In this paper we investigate applying LLMs to the problem of determining the directionality of edges in causal discovery. Specifically, we test our approach on a de-identified set of Non Small Cell Lung Cancer(NSCLC) patients that have both electronic health record and genomic panel data. Graphs are validated using Bayesian Dirichlet estimators using tabular data. Our result shows that LLMs can accurately predict the directionality of edges in causal graphs, outperforming existing state-of-the-art methods. These findings suggests that LLMs can play a significant role in advancing causal discovery and help us better understand complex systems.**

***Index Terms*—Large Language Models, Causal Discovery, Electronic Health Record, Genomics, Average Treatment Effect**

I. Introduction

Healthcare data analysis has been revolutionized in recent years with the application of Machine Learning (ML) and Deep Learning (DL) techniques [1]. But despite the success in predictive modelling, there is great interest in providing explainable models for the causal relationships between variables [2]. The current "black box" approach to modelling has limited interpretability and has not achieved acceptance in clinical settings [3], [4]. Causal modelling can provide an understanding of the underlying cause-effect relationship of the data, allowing counterfactual analysis to be performed [5]. In scenarios where cause-effect is not known a-priori Causal Discovery algorithms are used to identify causal relationships. State of the art Causal Discovery algorithms rely on score-based methods to generate Directed Acyclic Graphs (DAGs) but don't incorporate any domain expertise or expert knowledge.

The recent advancements in Large Language Models (LLMs) have sparked interest in their application for Causal Discovery or Causal Structure Learning (CSL) [6], [7]. LLMs have been proposed to serve as a surrogate for expert knowledge. Our method incorporates the feedback of LLM for CSL in the field of oncology [8]. We show the potential of applying LLMs in Causal Discovery by improving the accuracy of the generated Causal DAGs on data extracted from the Electronic Health Record (EHR) and Molecular Genomic Reports. Models are scored using the Bayesian Dirichlet equivalent uniform (Bdeu) score.

We demonstrate our approach on a dataset of Non Small Cell Lung Cancer (NSCLC) patients. In oncology molecular tests are already being used in the diagnosis to determine the course of treatment [9]. But there is a growing interest in finding new causative biomarkers and individualize treatments based on a patients medical history. Causal Modelling has been proposed as a key component in biomarker discovery [10].

II. Study Sample (Dataset)

We start off with a dataset of > 1000 deidentified patients extracted from Providence St. Joseph Health (PSJH's) clinical data warehouse. The cohort are the patients that have been diagnosed with NSCLC, a total of 455 patients. We only include patients that have a recorded smoking status, leaving a total of 326 patients remaining. The features used for this experiment consist of multi-modal data of Electronic Health

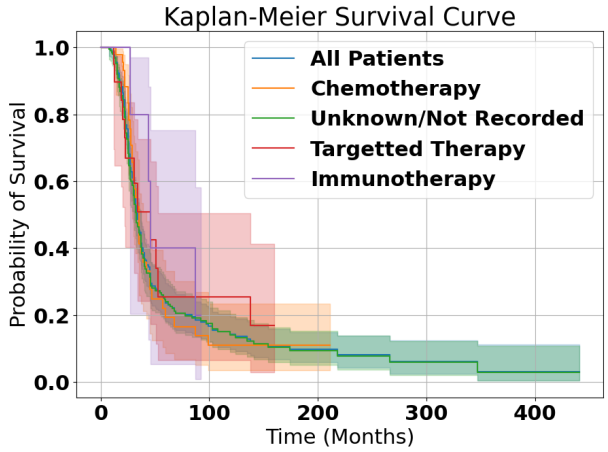

Fig. 1. Kaplan–Meier Survival Curve For Associated Treatment Plans

Record (EHR), and genomics. Genomics include the somatic mutation status of the genes KRAS, EGFR, FGFR1, ALK, MET, PIK3CA, BRAF, ROS1 and RET. From this set of only EGFR, ALK, ROS1, BRAF, NTRK, HER2, MET, RET, KRAS have an FDA approved therapy while PIK3CA, AKT1, PTEN are under clinical trials [11], [12]. The demographics of this data set is shown in Table I. The majority of the cases are Stage IV cancers and most patients are treated using standard chemotherapy. In addition, the majority of patients are non-smokers.

TABLE I
FEATURE SUMMARY

Characteristic	Summary
Number of Patients	326
Age	73.3±10.6
Survival Days	1179.7±1581.1
Sex	
Male	42.3%
Female	57.7%
Smoking Status	
Smoker	19.0%
Non-Smoker	81.0%
Stage	
I	30.4%
II	8.0%
III	14.1%
IV	47.5%
Treatment Plan	
Unknown/Not Recorded	69.3%
Chemotherapy	20.9%
Targeted Therapy	7.1%
Immunotherapy	2.8%
Genomics	
KRAS	27.9%
EGFR	39.0%
FGFR1	5.8%
ALK	20.6%
MET	11.0%
PIK3CA	26.4%
BRAF	5.2%
RET	42.9%

As shown in Figure 1 the Kaplan-Meier [13] shows that the survival probability for chemotherapy appear to decline over

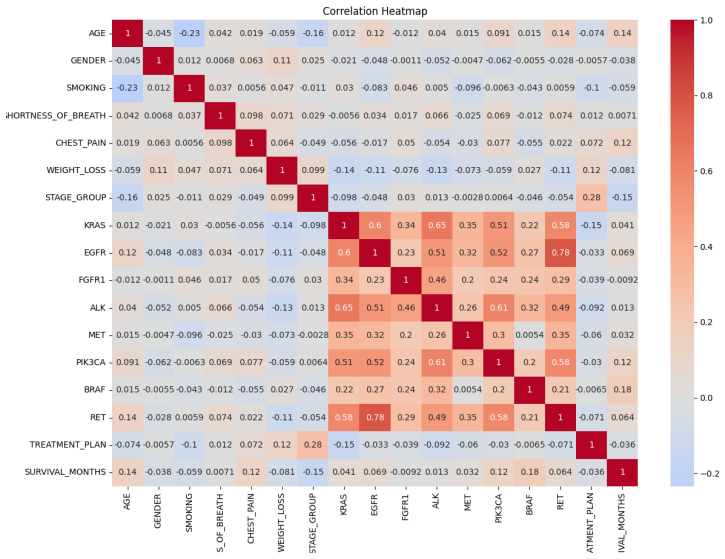

Fig. 2. Heatmap Showing Correlation of Features in the Dataset

time at a slower rate than the general population. The curve labelled "All Patients" illustrated the mean survival times of the entire patient cohort.

A heatmap showing the correlation between the chosen features is given in Figure 2. The correlation coefficient between any two variables is displayed in each cell. The hue of the cell indicates the direction and strength of the correlation: When the correlation is near to 1 , there is a strong positive correlation (the tendency for both variables to rise as one rises). If the correlation is near to -1, it is strongly negative (when one variable rises, the other tends to fall). Little or no association between the variables is indicated by a correlation that is near to 0. From this heatmap we see little correlation amongst the clinical variables but strong correlation among the genomic markers. Despite these correlation, few if any of these genomic markers are causally related. For example, both EGFR and KRAS are strongly associated with smokers in lung cancer. The more likely explanation is that both variables are being connected through a shared parent node [14].

III. METHODS

In this section we present our overall methods for DAG generation, selection of DAG and ATE for different treatment categories. This study was conducted under a protocol approved by the Providence IRB (Protocol 2018000188).

A. Causal Network Generation

While discovering causal structures from observational data is a difficult task, a number of strategies have been proposed. Most state of the art strategies rely on optimization techniques such as those used in the PC algorithm [15] and NOTEARS [16].

Instead we opt to start building the graph using LLMs to form the edges between the nodes. In this scenario the

LLM is being used as a surrogate for a domain expert to condition on the DAG. Initial experiments involved using zeroshot prompt [17] engineering to question the existing of an edge between each of the 18 feature nodes, requiring a total of $n(n-1)/2 = 153$ prompts. Each prompt was formatted to query a potential cause-effect relationship. The completions from the LLM are interpreted to determine if there is an edge in the causal graph in the specified direction. For the final model we request the LLM to reinvestigate how mutation is effecting the treatment plan and survival months. Finally, we request the LLM to update the model such that treatment plan should effect survival months. This final adjacency matrix represents the *final* version of the LLM model.

B. Graph Validation

A persistent problem when applying Causal Discovery in real world scenarios is that it is difficult to validate the generated graph. In addition due to the randomized nature of LLMs, even with low temperature settings response can vary. Both of these scenarios require some metric to determine which graph best fits the available data. Each of these DAGs represent a possible set of causal relationships between variables. Selecting the best or the optimal DAG among these becomes crucial in ensuring the reliable predictions and interpretations. Here we discuss different strategies to tackle this challenge.

1) Bayesian Estimation: In order to compare the "fit" of our proposed network we chose the Bayesian Dirichlet equivalent uniform ($Bdeu$) score [18]. The $Bdeu$ score is an extension of the Maximum Likelihood Probability that includes the prior probability of a given DAG and the marginal probability of the observed data. Mathematically it is written as the following:

$$Bdeu = logP(D|G) + logP(G) - logP(D) \tag{1}$$

For each node a separate $Bdeu$ score is calculated that best matches the Conditional Probability Distribution (CPD) of the parent nodes. When evaluating a singular node with a total number of N_j possible states and parent nodes with a total of N_i possible configurations the $Bdeu$ score simplifies to the following equation:

$$Bdeu = \sum_{i=1}^{N_i} \sum_{j=1}^{N_j} \left(n_{ij} + \frac{\alpha}{N_j} \right) \log \left(\frac{n_{ij} + \frac{\alpha}{N_j}}{n_i + \alpha} \right) \tag{2}$$

where n_ij is the number of times the node takes on its j-th class given that the parent nodes are in their i-th configuration. n_i is the total number of times the input nodes are in their i-th configuration and α is a heuristic parameters which is often used to represent the Equivalent Sample Size (ESS). Intuitively the $Bdeu$ score is often chosen due it's ability to penalize complex models, particular models with large numbers of parent nodes [19]. At the same time it does not require prior domain knowledge and is computationally efficient [20]. The total $Bdeu$ score of a model is simply the sum of all $Bdeu$ scores for each node.

IV. Interpretability and Results

A key advantage of using causal modelling is that they ensure trust in research and healthcare policies, emphasise the effects of interventions and make sure decisions are based on accurate evidence [21]. In this section we will discuss the DAGs that can be generated using the the LLM responses. The main method used for interpreting and validation is by modelling these DAGs as Bayesian networks. Bayesian networks are natural models for Causal DAGs, and have been shown to effective at inferring and evaluating outcomes of causal inference [22]. Fitting and evaluating the Conditional Probability Distributions for each node in the network is done using the Bdeu score [19]. For measuring the intervention effect, we calculate the Average Treatment Effect (ATE) given both mutation status and treatment combination [23].

A. DAG Generation

For the single prompt method, we utilized OpenAI's Codeinterpreter plugin for GPT-4 that parses the input prompt into a singlular edge adjacency matrix. The DAG showing our final model is given in Figure 3.

To compare our methods to existing causal discover algorithms, we applied the NOTEARS and PC algorithms to our data set. The NOTEARS algorithm learns a graph structure by minimizing a continuous, differentiable objective function [16]. The PC algorithm instead starts with a fully connected graph and performs statistical independence tests to removes edges. After removing non-statistically dependence nodes, the algorithm enforces acyclicity on the remaining edges using a set of four rules [24]. The NOTEARS implementation was done using the *causalnex* library while the PC algorithm used the *gcastle* library [25], [26]. The DAGs generated from NOTEARS, PC algorithm, etc are found in the Appendix.

B. Validation

Each DAG is modelled as a Bayesian Network and fit to the data set. This Bayesian approach allows us to make probabilistic statements about the causal effect β based on both our prior beliefs and the observed data. It naturally incorporates uncertainty and provides a richer understanding than a single point estimate. For each model, the adjacency matrix was used to create a Bayesian Network. The Conditional Probability Distributions (CPD) were fit for each node using the Bdeu score as a criteria. Creating and optimizing the Bayesian network along with calculating the final Bdeu score was done using the *pgmpy* Python package [27].

For each of the Bayesian networks generated, the Bdeu score was calculated using the observational NSCLC dataset to compare how each model fit to the observed data. The score for the five models are given in Table III. We note that the LLM methods vastly outperform optimization based approaches. The best score achieved through LLMs were -4150 while the NOTEARS and PC algorithm was only about to achieve -6886 and -6092 respectively.

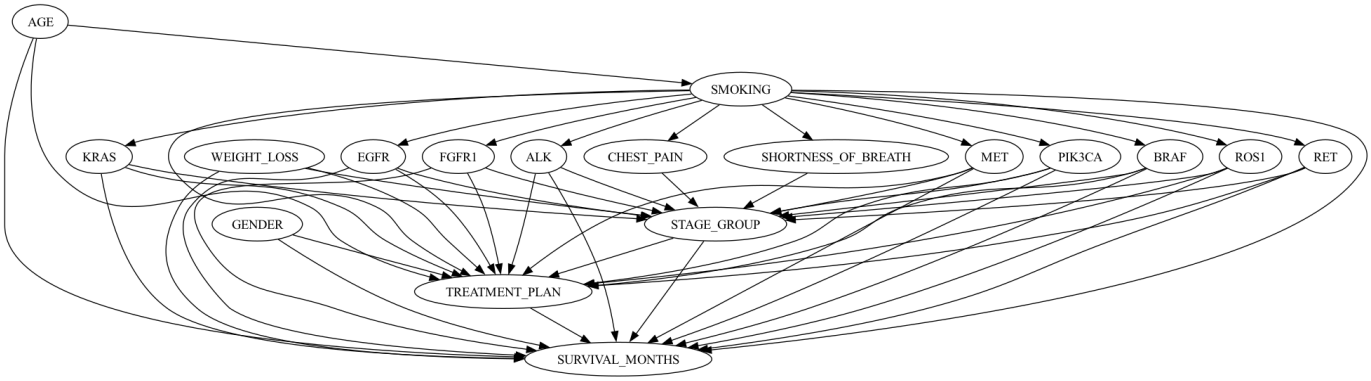

Fig. 3. DAG Generated using LLM.

TABLE II
ATE WITH DIFFERENT TREATMENT CATEGORY AND MUTATION EVIDENCE

Treatment Category	KRAS	EGFR	FGFR1	ALK	MET	PIK3CA	BRAF	RET
Chemotherapy	0.028753	0.027040	0.024604	0.027979	0.025711	0.028071	0.024134	0.030823
Targeted Therapy	0.020828	0.021891	0.015868	0.018419	0.016813	0.019083	0.015762	0.023738
Immunotherapy	0.007267	0.003912	0.005502	0.006562	0.005807	0.006607	0.005356	0.004674

TABLE III
BDEU SCORE FOR THE LLM GENERATED CAUSAL GRAPH FITTED TO BAYESIAN NETWORKS

Equivalent Sample Size	LLM	NOTEARS	PC
5	-4228	-7037	-6418
10	-4171	-6935	-6202
15	-4150	-6886	-6092

C. Interpreting

The ATE is calculated using the variable elimination method to compute the conditional probability of survival based on getting the treatment and having the mutation.

$$ATE = E[Y_1 - Y_0] \tag{3}$$

where E is the expectation, Y_1 is treated and Y_0 is controlled group. The ATE values for various treatments depending on certain gene mutations are shown in Table II. The ATE on patients with a specific gene mutation compared to people without it. Different gene mutations are shown by KRAS, EGFR, and ALK columns, while chemotherapy and immunotherapy are indicated by rows. The information aids in understanding each treatment's efficacy for people with a particular mutation and directs individualised medical choices. We observe that the presence of a RET mutation has the largest effect for both Chemotherapy and Targeted Therapies. In the targeted therapy category the FGFR1 biomarker displaying the least interaction at 0.015868. Notably, the immunotherapy treatment category had poor scores across all biomarkers. This is not too surprising as the most effective markers for Immunotherapy are PD-L1 expression and Tumor Mutational burden, neither of which are represented in the current set of markers [28]. These results suggest varying degrees of influence of these treatment modalities on different biomarkers.

V. CONCLUSION & FUTURE WORK

There is great potential for LLMs to revolutionize healthcare. We have investigated how LLMs can be applied in the generation of Causal DAGs. Using the Bdeu score as a metric, we evaluated that LLM methods outperformed existing Causal Discovery methods.

Further data modalities will improve the accuracy of a patient's response to therapy. The intricacy of cause effect analysis for NSCLC may not be fully captured by this restricted dataset, which could lead to an oversimplification of complicated causal linkages. Future work would expand the amount of genomic variations being examined and include advanced genomic markers. Including other histopathological biomarkers, such as PD-L1, will also improve accuracy of the model [29].

The final limitation we observed is the use of general purpose LLM models. These LLMs have been trained on a wide variety of tasks and lack the specialized training to determine causal relationships in the medical domain. Future versions will incorporate models that have been specially trained on medical literature such as Med-PaLM2 or BioGPT [30]–[32].

REFERENCES

[1] S. Yang, F. Zhu, X. Ling, Q. Liu, and P. Zhao, "Intelligent health care: Applications of deep learning in computational medicine," vol. 12. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fgene.2021.607471

[2] J. Shi and B. Norgeot, "Learning causal effects from observational data in healthcare: A review and summary," vol. 9. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fmed.2022.864882

[3] A. Chaddad, J. Peng, J. Xu, and A. Bouridane, "Survey of explainable AI techniques in healthcare," vol. 23, no. 2, p. 634. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9862413/

[4] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A review of machine learning interpretability methods," vol. 23, no. 1, p. 18, number: 1 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/1099-4300/23/1/18

[5] J. PEARL, "Causal diagrams for empirical research," *Biometrika*, vol. 82, no. 4, pp. 669–688, 12 1995. [Online]. Available: https://doi.org/10.1093/biomet/82.4.669

[6] C. Zhang, S. Bauer, P. Bennett, J. Gao, W. Gong, A. Hilmkil, J. Jennings, C. Ma, T. Minka, N. Pawlowski *et al.*, "Understanding causality with large language models: Feasibility and opportunities," *arXiv preprint arXiv:2304.05524*, 2023.

[7] X. Liu, D. Yin, C. Zhang, Y. Feng, and D. Zhao, "The magic of IF: Investigating causal reasoning abilities in large language models of code," in *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics, pp. 9009–9022. [Online]. Available: https://aclanthology.org/2023.findings-acl.574

[8] D. D. Shapiro and P. Msaouel, "Causal diagram techniques for urologic oncology research," *Clinical Genitourinary Cancer*, vol. 19, no. 3, pp. 271.e1–271.e7, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1558767320301932

[9] R. Pirker, "Conquering lung cancer: current status and prospects for the future," *Pulmonology*, vol. 26, no. 5, pp. 283–290, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2531043720300313

[10] C. Cai, G. F. Cooper, K. N. Lu, X. Ma, S. Xu, Z. Zhao, X. Chen, Y. Xue, A. V. Lee, N. Clark, V. Chen, S. Lu, L. Chen, L. Yu, H. S. Hochheiser, X. Jiang, Q. J. Wang, and X. Lu, "Systematic discovery of the functional impact of somatic genome alterations in individual tumors through tumor-specific causal inference," *PLoS Computational Biology*, vol. 15, no. 7, p. e1007088, Jul. 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6650088/

[11] A. C. Tan, "Targeting the PI3k/akt/mTOR pathway in non-small cell lung cancer (NSCLC)," vol. 11, no. 3, pp. 511–518.

[12] Z. Sirhan, R. Alojair, A. Thyagarajan, and R. P. Sahu, "Therapeutic implications of PTEN in non-small cell lung cancer," vol. 15, no. 8, p. 2090, number: 8 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/1999-4923/15/8/2090

[13] M. K. Goel, P. Khanna, and J. Kishore, "Understanding survival analysis: Kaplan-meier estimate," *International journal of Ayurveda research*, vol. 1, no. 4, pp. 274–278, 2010.

[14] K. TAKAMOCHI, S. OH, and K. SUZUKI, "Differences in EGFR and KRAS mutation spectra in lung adenocarcinoma of never and heavy smokers," vol. 6, no. 5, pp. 1207–1212. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3813793/

[15] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search*, ser. Lecture Notes in Statistics, J. Berger, S. Fienberg, J. Gani, K. Krickeberg, I. Olkin, and B. Singer, Eds. New York, NY: Springer, 1993, vol. 81. [Online]. Available: http://link.springer.com/10.1007/978-1-4612-2748-9

[16] X. Zheng, B. Aragam, P. K. Ravikumar, and E. P. Xing, "DAGs with NO TEARS: Continuous Optimization for Structure Learning," in *Advances in Neural Information Processing Systems*, vol. 31. Curran Associates, Inc., 2018.

[17] M. Suzgun, L. Melas-Kyriazi, and D. Jurafsky, "Prompt-and-rerank: A method for zero-shot and few-shot arbitrary textual style transfer with small language models," *arXiv preprint arXiv:2205.11503*, 2022.

[18] D. Heckerman, D. Geiger, and D. M. Chickering, "Learning bayesian networks: The combination of knowledge and statistical data," vol. 20, no. 3, pp. 197–243, company: Springer Distributor: Springer Institution: Springer Label: Springer Number: 3 Publisher: Kluwer Academic Publishers-Plenum Publishers. [Online]. Available: https://link.springer.com/article/10.1023/A:1022623210503

[19] Z. Liu, B. Malone, and C. Yuan, "Empirical evaluation of scoring functions for bayesian network model selection," vol. 13, no. 15, p. S14. [Online]. Available: https://doi.org/10.1186/1471-2105-13-S15-S14

[20] M. Scutari, "Dirichlet bayesian network scores and the maximum relative entropy principle," vol. 45, no. 2, pp. 337–362. [Online]. Available: https://doi.org/10.1007/s41237-018-0048-x

[21] N. Almeda, C. R. García-Alonso, J. A. Salinas-Pérez, M. R. Gutiérrez-Colosía, and L. Salvador-Carulla, "Causal Modelling for Supporting Planning and Management of Mental Health Services and Systems: A Systematic Review," *International Journal of Environmental Research and Public Health*, vol. 16, no. 3, p. 332, Feb. 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6388254/

[22] J. Pearl, "From Bayesian Networks to Causal Networks," in *Mathematical Models for Handling Partial Knowledge in Artificial Intelligence*, G. Coletti, D. Dubois, and R. Scozzafava, Eds. Boston, MA: Springer US, 1995, pp. 157–182.

[23] A. I. Naimi and B. W. Whitcomb, "Defining and Identifying Average Treatment Effects," *American Journal of Epidemiology*, vol. 192, no. 5, pp. 685–687, 01 2023. [Online]. Available: https://doi.org/10.1093/aje/kwad012

[24] J. Zhang, "On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias," *Artificial Intelligence*, vol. 172, no. 16, pp. 1873–1896, Nov. 2008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0004370208001008

[25] P. Beaumont, B. Horsburgh, P. Pilgerstorfer, A. Droth, R. Oentaryo, S. Ler, H. Nguyen, G. A. Ferreira, Z. Patel, and W. Leong, "CausalNex," Oct. 2021. [Online]. Available: https://github.com/quantumblacklabs/causalnex

[26] K. Zhang, S. Zhu, M. Kalander, I. Ng, J. Ye, Z. Chen, and L. Pan, "gCastle: A Python Toolbox for Causal Discovery," Nov. 2021, arXiv:2111.15155 [cs, stat]. [Online]. Available: http://arxiv.org/abs/2111.15155

[27] A. Ankan and A. Panda, "pgmpy: Probabilistic graphical models using python," in *Proceedings of the 14th Python in Science Conference (SCIPY 2015)*. Citeseer, 2015.

[28] M. Yarchoan, L. A. Albacker, A. C. Hopkins, M. Montesion, K. Murugesan, T. T. Vithayathil, N. Zaidi, N. S. Azad, D. A. Laheru, G. M. Frampton, and E. M. Jaffee, "PD-L1 expression and tumor mutational burden are independent biomarkers in most cancers," *JCI Insight*, vol. 4, no. 6, p. e126908. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6482991/

[29] P. BULUTAY, P. FIRAT, H. ZEREN, S. ERUS, S. TANJU, and DİLEGE, "The importance of histological patterns on PD-L1 staining heterogeneity: Should we use pattern-based approach for selecting tumor samples for PD-L1 testing in lung adenocarcinomas?" *Turkish Journal of Medical Sciences*, vol. 51, no. 1, pp. 204–213, Feb. 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7991888/

[30] K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, L. Hou, K. Clark, S. Pfohl, H. Cole-Lewis, D. Neal *et al.*, "Towards expert-level medical question answering with large language models," *arXiv preprint arXiv:2305.09617*, 2023.

[31] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, P. Payne, M. Seneviratne, P. Gamble, C. Kelly, A. Babiker, N. Schärli, A. Chowdhery, P. Mansfield, D. Demner-Fushman, B. Agüera y Arcas, D. Webster, G. S. Corrado, Y. Matias, K. Chou, J. Gottweis, N. Tomasev, Y. Liu, A. Rajkomar, J. Barral, C. Semturs, A. Karthikesalingam, and V. Natarajan, "Large language models encode clinical knowledge," vol. 620, no. 7972, pp. 172–180, number: 7972 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41586-023-06291-2

[32] R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, and T.-Y. Liu, "BioGPT: generative pre-trained transformer for biomedical text generation and mining," vol. 23, no. 6, p. bbac409. [Online]. Available: https://doi.org/10.1093/bib/bbac409

APPENDIX

To generate an DAG using an LLM, prompts where provided that could determine causal linkage and directionality for each node. Each prompt was formatted to query a potential cause-effect relationship. Afterwards the completions from the LLM are interpreted to determine if there is an edge in the causal graph in the specified direction. Table IV provides a list of example prompts, the associated causal edges that they test, the response and the final assumption for the causal test. Initial experiments involved using zeroshot prompt engineering to question the existing an edge between each of the 18 feature nodes, requiring a total of $n(n-1)/2 = 153$ prompts

For the next model we used a single prompt that contains all the nodes and prompt the LLM to generate the full graph. The full prompt and output is shown in Table V.

Afterwards, we started updating the model using human intervention. We first corrected the lack of AGE influencing smoking status of a patient with the following prompt, "how age is not cause smoking please relook into the adjacency matrix and generate a correct one." Then we add in that staging and smoking should be effecting the mutation status, "the stage group and smoking should cause some mutation in nsclc." This represents the *V3* LLM DAG.

For the final model we request the LLM to re-evaluate how mutation is effecting the treatment plan and survival months. Finally, we request the LLM to update the model such that treatment plan should effect survival months. This final adjacency matrix represents the final version of the DAG generated by the LLM.

Figure 6 shows the DAG generated based on NOTEARS algorithm and Figure 7 shows DAG with PC algorithm. It should be noted that both methods are unable to generate clinical relevance DAGs using the provided dataset. For example, in the NOTEARS algorithm DAG we have several symptoms causing AGE in addition to a view somatic mutations. The generated DAGs based on the observed data with PC and NOTEARS most of the edges are not oriented(i.e direction) which is essential for treatment effect analysis.

It should be noted that both PC and NOTEARS are unable to generate clinical relevant DAGs from the dataset. For example, in the NOTEARS' DAG we have several symptoms causing AGE in addition to somatic mutations.

TABLE IV
FROM LEFT TO RIGHT. AN EXAMPLE PROMPT USED TO TEST FOR THE PROVIDE EDGE, THE CORRESPONDING COMPLETION AND THE FINAL VERDICT.

Prompt	Edges	Completion/Response	Assumption
Does age effect Gender in NSCLC	Age→Gender	Age and gender can both independently influence the development and characteristics of non-small cell lung cancer (NSCLC), but they do not have a direct cause-and-effect relationship with each other	No
Does age effect survival analysis for NSCLC	Age→Survival	Yes, age can have an impact on survival analysis	Yes
Does age causes shortness of breathe in NSCLC	Age→Shortness of breath	Age itself does not directly cause shortness of breath in NSCLC	No
Does KRAS mutation effect survival in NSCLC	KRAS→Survival	KRAS mutation subtype: There are different subtypes of KRAS mutations and some studies have suggested that specific subtypes may have varying impacts on survival	Yes
Does targeted therapy effect survival in NSCLC	Treatment Plan→Survival	Therapy can have a significant impact on the survival outcomes of patients with NSCLC, particularly those with specific molecular alterations that are targeted by the therapy.	Yes

TABLE V
SINGLE PROMPT INPUT AND RESPONSE

Prompt	Generate me a cause effect adjacency matrix for these nodes AGE, SMOKING, GENDER, SHORTNESSOFBREATH, CHESTPAIN, WEIGHTLOSS, TREATMENTPLAN, SURVIVALMONTHS, STAGEGROUP, KRAS, EGFR, FGFR1, ALK, MET, PIK3CA, BRAF, ROS1, RET mutation doesn't cause symptoms.
Response	In order to create a cause-effect adjacency matrix, we need to understand the relationships between the given nodes. Here's a possible interpretation of the relationships between them: AGE can affect the TREATMENTPLAN and SURVIVALMONTHS. SMOKING can lead to CHESTPAIN, SHORTNESSOFBREATH, and can affect the TREATMENTPLAN, SURVIVALMONTHS, and STAGEGROUP. GENDER can affect the TREATMENTPLAN and SURVIVALMONTHS. SHORTNESSOFBREATH and CHESTPAIN can indicate the STAGEGROUP, which in turn influences the TREATMENTPLAN and SURVIVALMONTHS. WEIGHTLOSS can also indicate the STAGEGROUP and can affect the TREATMENTPLAN and SURVIVAL_MONTHS. Mutations (KRAS, EGFR, FGFR1, ALK, MET, PIK3CA, BRAF, ROS1, RET) do not cause symptoms (as per the user's instructions) but they can affect the TREATMENTPLAN, SURVIVALMONTHS, and STAGEGROUP.

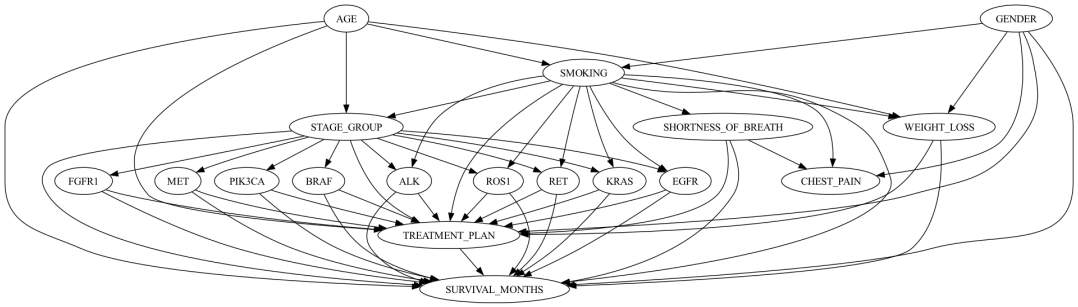

Fig. 4. Causal DAG LLM sequence prompts and adjacency matrix

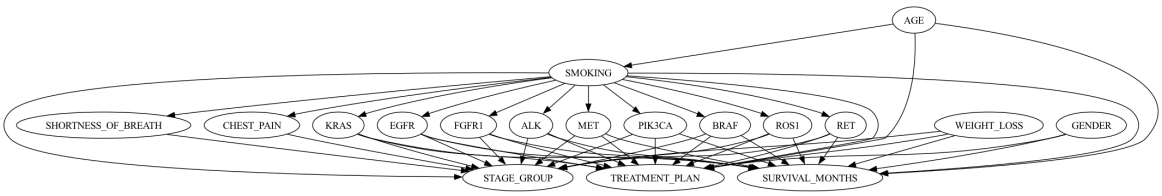

Fig. 5. DAG Generated for the *V3* version of LLM Model

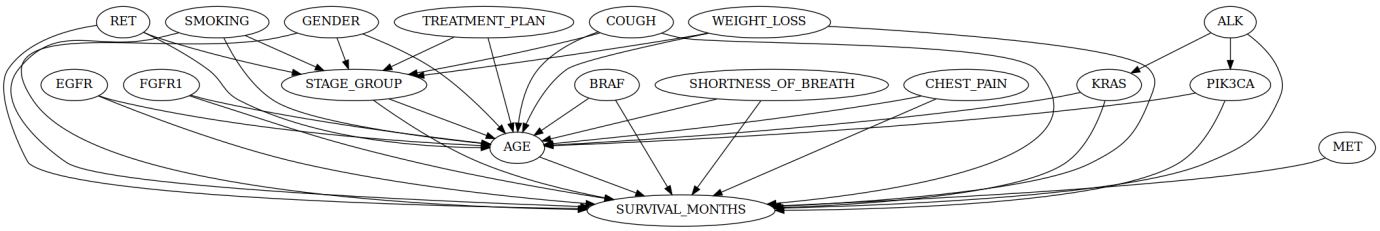

Fig. 6. Causal DAG generated with NoTears, $max_{iter} = 100$, $h_{tol} = 1e-08$, $w_{threshold} = 0.5$

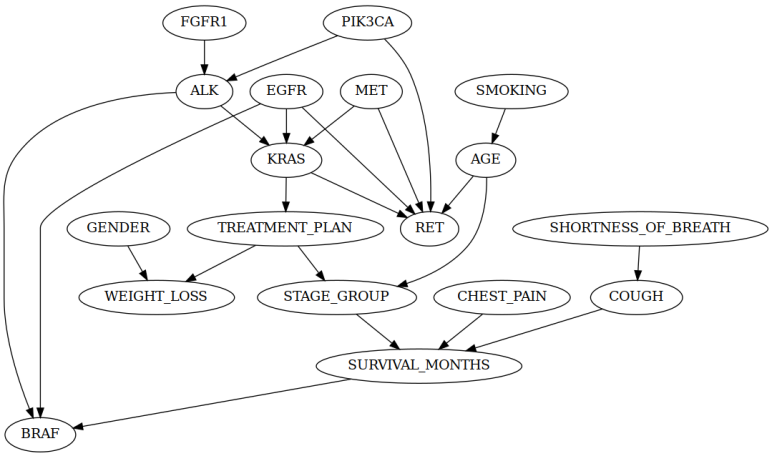

Fig. 7. DAG Generated using PC Algorithm