RESEARCH

Solving the puzzle of quality of life in cancer: integrating causal inference and machine learning for data-driven insights

Hakan Şat Bozcuk[1*] and Mustafa Serkan Alemdar[1]

Abstract

Background Understanding the determinants of global quality of life in cancer patients is crucial for improving their overall well-being. While correlations between various factors and quality of life have been established, the causal relationships remain largely unexplored. This study aimed to identify the causal factors influencing global quality of life in cancer patients and compare them with known correlative factors.

Methods We conducted a retrospective analysis of European Organization for Research and Treatment of Cancer Quality of Life Questionnaire data, alongside demographic and disease-related features, collected from new cancer patients during their initial visit to an oncology outpatient clinic. Correlations with global quality of life were identified using univariate and multivariate regression analyses. Causal inference analysis was performed using two approaches. First, we employed the Dowhy Python library for causal analysis, incorporating prior information and manual characterization of an acyclic graph. Second, we utilized the Linear Non-Gaussian Acyclic Model (LiNGAM) machine learning algorithm from the Lingam Python library, which automatically generated an acyclic graph without prior information. The significance level was set at $p < 0.05$.

Results Multivariate analysis of 469 new admissions revealed that disease stage, role functioning, emotional functioning, social functioning, fatigue, pain and diarrhea were linked with global quality of life. The most influential direct causal factors were emotional functioning, social functioning, and physical functioning, while the most influential indirect factors were physical functioning, emotional functioning, and fatigue. Additionally, the most prominent total causal factors were identified as type of cancer (diagnosis), cancer stage, and sex, with total causal effect ratios of -9.47, -4.67, and − 1.48, respectively. The LiNGAM algorithm identified type of cancer (diagnosis), nausea and vomiting and social functioning as significant, with total causal effect ratios of -9.47, -0.42, and 0.42, respectively.

Conclusions This study identified that causal factors for global quality of life in new cancer patients are distinct from correlative factors. Understanding these causal relationships could provide valuable insights into the complex dynamics of quality of life in cancer patients and guide targeted interventions to improve their well-being.

Keywords Causal inference, Quality of life, Cancer, Machine learning, Regression analysis, Social functioning, Emotional functioning

*Correspondence:
Hakan Şat Bozcuk
hbozcuk@gmail.com
[1]Dept. of Medical Oncology, Medical Park Hospital, Antalya, Turkey

Background
In recent decades, while the primary focus of cancer treatment has been on achieving a cure or prolonging survival, there has been a growing recognition of the importance of improving the quality of life (QoL) for patients with cancer [1]. Various scales and questionnaires are used to assess QoL in different cancer types, such as breast cancer, but not all of these tools have been universally validated across all settings and cultures [2, 3]. One of the most extensively studied instruments for cancer patients is the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire (EORTC QLQ-C30) [4, 5]. Although global QoL is a key dimension of this scale, it is not merely the sum of its individual parts, which encompass various dimensions of QoL [6]. Despite our understanding that many disease- or treatment-related factors may influence specific QoL dimensions, the underlying causal factors of global QoL remain largely unexplored.

Although predictive and explanatory models in the field of medicine have predominantly adopted a correlational approach, it is essential to recognize that correlation does not imply causation. Identifying causal factors is crucial for evaluating the impact of interventions, policy changes, or exposure to certain factors on outcomes, particularly in the medical domain [7]. Artificial intelligence (AI) has found widespread applications in medicine for tasks such as diagnosis, research, and prediction, often leveraging correlated data [8]. However, to overcome the limitations of working solely with correlated data, it is imperative to conduct special analyses, such as causal inference. This kind of analysis theoretically enables clinicians and researchers to discern causal factors and evaluate treatment effects using observational data, thereby providing insights into complex networks of relationships [9]. Therefore, gaining a deeper understanding of the causal factors underlying QoL is pivotal for enhancing the overall well-being and QoL of cancer patients.

Machine learning (ML), a critical subset of artificial intelligence (AI), involves the creation of algorithms that can learn from data and make predictions or decisions without explicit programming for specific tasks. This capability allows ML to analyze vast and complex datasets, identifying patterns and making informed predictions that can significantly enhance various domains, including medicine. Integrating ML with causal inference allows researchers and clinicians to not only predict outcomes but also understand the underlying mechanisms driving these outcomes, leading to more effective and targeted interventions.

For this reason, in this study, we sought to explore the causal factors influencing global QoL by analyzing observational data collected from consecutive cancer patients during their initial appointments at a cancer center. By applying analytical methods capable of identifying causal relationships, we aimed to shed light on the intricate web of factors influencing global QoL in cancer patients.

Methods
Patient selection and data collection
All patients with cancer who visited a private hospital's cancer center and were under the care of the same attending physician were provided with the Turkish version of the EORTC QLQ-C30 version 3 questionnaire during their initial appointment at the center [10]. We included quality of life data from all consecutive new patients who consented to complete the questionnaire between March 2018 and April 2021. Additionally, we collected primary demographic and disease-related information.

Quality-of-life dimension scores were computed following the guidelines outlined in the scoring manual. Quality of life dimensions of the EORTC QLQ-C30 questionnaire were QL2; global quality of life, PF2; physical functioning, RF2; role functioning, EF; emotional functioning, CF; cognitive functioning, SF; social functioning, FA; fatigue, NV; nausea and vomiting, PA; pain, DY; dyspnea, SL; insomnia, AP; appetite loss, CO; constipation, DI; diarrhea, and FI; financial difficulties. The study was approved by the ethical committee of the Medical Park Hospital, Antalya, on the 17th of January 2024, with a reference number of 2024/1.

Evaluation of correlative factors for global quality of life
In order to contrast with the findings from evaluation of causal factors for global quality of life, we first wanted to conduct a non-causal, univariate regression analysis of potential predictor variables for global QoL (QL2). Variables with a p value of 0.10 or less in the univariate analysis were then included in the multivariate regression analysis. Stepwise method was then used in the multivariate analysis to select the significant variables. For diagnostic purposes, we categorized patients with breast cancer versus other cancers or with lung or colorectal cancer versus other cancers. We considered a p value less than 0.05 to indicate statistical significance. The statistical analyses were conducted using SPSS version 21.

Evaluation of causal factors for global quality of life
To investigate causal inference, we employed two distinct approaches. Firstly, we manually constructed a directed acyclic graph (DAG) using domain knowledge and the results of the preliminary exploratory analysis of the dataset (Expert-driven construction of the DAG). DAGs model the causal relationships among the variables of interest. DAGs are graphical representations where nodes signify variables and directed edges denote causal effects, with the constraint that no cycles exist, ensuring

a unidirectional flow of causation. We constructed our DAGs based on domain knowledge and prior empirical evidence to accurately capture the underlying causal mechanisms. These graphs were instrumental in identifying confounding variables, mediators, and potential sources of bias, thus guiding our statistical analysis and ensuring the robustness of our causal inferences. By leveraging DAGs, we aimed to achieve a clear and precise delineation of the causal pathways influencing the outcomes of our study, ultimately enhancing the validity and interpretability of our findings.

To construct DAG, we utilized DoWhy Python library, which is used for evaluating causal inference, to calculate and plot the separate impacts of direct, indirect, and total causal effects [11, 12]. Percentage direct causal strength quantifies the influence of a specific factor (or node) on another within a directed acyclic graph (DAG) in terms of direct causation. It measures the proportion of total direct causal impact exerted by each factor relative to the sum of all direct causal impacts on the target variable. Additionally, the percentage indirect causal effect quantifies the influence of a specific factor on an outcome through one or more intermediary variables within a DAG. It measures the proportion of the total indirect causal impact exerted by each factor relative to the sum of all indirect causal impacts on the target variable. As stated above, to be included in DAG, factors either had to be associated with each other during the exploratory statistical analysis, or had to be linked according to the oncology domain knowledge.

Secondly, we utilized the Lingam Python library to implement a fully automated machine learning approach based on the Linear Non-Gaussian Acyclic Model (LiN-GAM) [13]. LiNGAM is a machine learning algorithm used in causal discovery, specifically designed to identify causal structures from observational data. LiNGAM assumes that the data follows a linear relationship and that the errors (or noise) are non-Gaussian. This assumption allows the algorithm to distinguish between cause and effect, which is a significant advantage over other linear methods that might not be able to differentiate between the two. LiNGAM is also used in fields like econometrics, genomics, and various social sciences where understanding the causal relationships between variables is crucial. It operates under the principle that if the data generation process follows a linear non-Gaussian model, then it is possible to identify the causal ordering of the variables and the corresponding causal effects. In our study, LiNGAM allowed us to visualize the LiN-GAM adjacency matrix and calculate the total causal effects of various factors [14, 15]. We employed these two approaches (DAG and LiNGAM) basically to more soundly and validly test causal assumptions and relationships that are hidden in our dataset.

These analyses aimed to uncover not only the associational relationships between variables and global quality of life but also the potential causal factors influencing it. By employing the above statistical methods and leveraging both domain knowledge and automated machine learning algorithms, we sought to gain a deeper understanding of the complex interplay of factors shaping the overall quality of life experienced by cancer patients attending the center in the specified time period.

During the preparation of this work, the authors used Chat GPT 3.5 to improve readability. After using this tool, the authors reviewed and edited the content.

Results

General results
A total of 469 patients were recruited, with a median age of 56 years and a 63% female distribution. Breast cancer was the most common diagnosis and was noted in 40% of patients. The other cancers category (37%) included gastric cancer (4%), prostate cancer (4%), testis cancer (4%), skin cancer (2%) and ovarian cancer (2%) as the most prominent subtypes, and remaining various other cancer types (21%). The patient demographics and details can be found in Table 1. The median value for QL2 was 58 on a scale of 100. The distribution of QL2 is presented in Fig. 1.

Associates of global quality of life
According to the univariate analysis, all quality-of-life dimensions individually and diagnosis (lung or colorectal vs. other cancers) and disease stage were associated with the QL2 score. The feature with the highest t value was fatigue (FA) scores, with a beta of -0.61, t = -16.60, and $P < 0.001$. According to the multivariate analyses, 7 factors were significant, namely, disease stage, role functioning (RF2), emotional functioning (EF), social functioning (SF), fatigue (FA), pain (PA) and diarrhea (DI), and the factor with the highest t value was SF, with a beta of 0.18, t = 3.71, and $P < 0.001$. The details of the regression analyses can be found in Table 2.

Causal factors for global quality of life
Expert-driven construction of the DAG
With expert-driven construction of the DAG, the resultant model was found to provide explanatory value among the possible permuted DAGs, as shown by a *P* value LMC = 0.02 and a *P* value TPa < 0.01, which are both statistically significant. The plot for causal model evaluation is given in Fig. 2. The plot and the accompanying tests show that the given DAG structure has a lower fraction of violations for both the Local Markov Condition and Transitive Pairing compared to randomly permuted DAGs. The low p-values suggest that these results are statistically significant, indicating that the DAG is a

Table 1 Demographics and quality of life data

Feature	n	%	min-max	median	mean
Clinical					
Total	469	100			
Age			18-88	56	53
Sex					
Male	175	37			
Female	294	63			
Ecog status			0-4	1	0.8
Diagnosis					
Breast cancer	189	40			
Lung or colorectal cancer	108	23			
Other cancers	172	37			
Stage			1-4	3	2.6
Active treatment					
Yes	262	56			
No	207	44			
Time since diagnosis					
Less than 6 months	360	77			
6 to 12 months	31	7			
More than 12 months	78	16			
Quality of life related					
QL2			0-100	58	57
PF2			0-100	80	74
RF2			0-100	83	76
EF			0-100	75	73
CF			0-100	100	86
SF			0-100	83	78
FA			0-100	33	35
NV			0-100	0	8
PA			0-100	17	27
DY			0-100	0	16
SL			0-100	33	29
AP			0-100	0	19
CO			0-100	0	17
DI			0-100	0	11
FI			0-100	0	23

Quality of life dimensions: QL2; global quality of life, PF2; physical functioning, RF2; role functioning, EF; emotional functioning, CF; cognitive functioning, SF; social functioning, FA; fatigue, NV; nausea and vomiting, PA; pain, DY; dyspnea, SL; insomnia, AP; appetite loss, CO; constipation, DI; diarrhea, FI; financial difficulties.
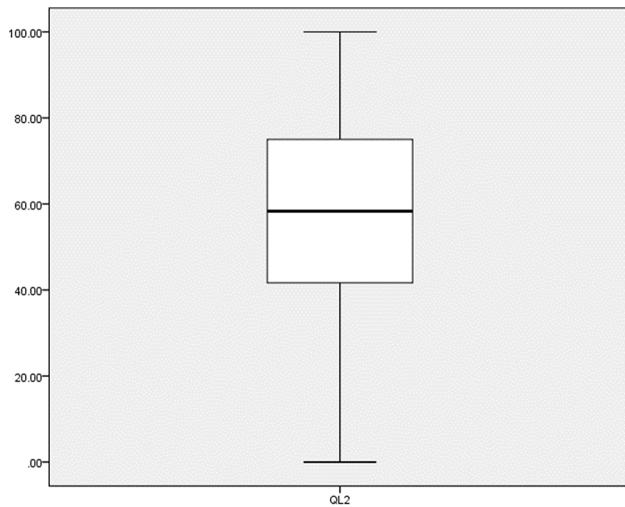
Fig. 1 Distribution of Global Quality of Life (QL2) scores. Global quality of life distribution. QL2; Global quality of life. The Y-axis shows the scores on a scale ranging from 0 to 100

valid and explanatory model for the underlying causal relationships in the data.

The plotted DAG with the network of causal relationships among the causal factors and QL2 is detailed and can be visualized in Fig. 3.

The DAG structure has some underlying patterns, although not all causal effects are of large amplitude, for example, a number of clinical factors and quality of life domains on the left column in Fig. 3 appear to effect PF2; physical functioning, among others, and PF2 possesses direct and indirect causal effects (through EF; emotional functioning and CF; cognitive functioning) on QL2; global quality of life.

Among the direct causal factors, the most influential was emotional functioning, with a percentage direct causal strength of 38%. When indirect causal factors were considered, physical functioning was at the top, with a percentage indirect causal strength of 30%. Finally, among the total causal factors, diagnosis (lung or colorectal cancer versus other cancers) was the most influential causal factor, with a causal effect ratio of -9.47, other prominent total causal factors were stage and sex. DAG did not show valid paths for some of the causal associations between various factors, including diagnosis, and QL2, which is the main outcome of interest. In causal analysis, unobserved variables (also called latent variables) can play a crucial role in the causal relationships between observed variables, and as such, the plotted DAG we formulated may conceal some of these latent variables. The plots for causal factors found by the DAG

Table 2 Associates of global quality of life

Feature	Univariate analyis			Multivariate analyis		
	Beta*	t	P	Beta*	t	P
Clinical						
Age	-0.04	-0.95	0.341			
Sex (male vs. female)	-0.03	-0.60	0.551			
Diagnosis-1*** (breast vs. other cancers)	0.00	-0.00	0.998			
Diagnosis-2*** (lung or colorectal vs. other cancers)	-0.15	-3.00	**0.003**	-0.04	-1.07	0.287
Stage	-0.22	-4.76	**<0.001**	-0.08	-2.15	**0.032****
Active treatment (yes vs. no)	-0.02	-0.50	0.617			
Time since diagnosis (less than 6 vs. 6 to 12 vs. more than 12 months)	-0.01	-0.14	0.893			
Quality of life related						
PF2	0.50	12.30	**<0.001**	0.08	1.53	0.128
RF2	0.52	13.30	**<0.001**	0.13	2.65	**0.008****
EF	0.40	9.33	**<0.001**	0.15	3.70	**<0.001****
CF	0.33	7.57	**<0.001**	-0.02	-0.37	0.709
SF	0.54	13.92	**<0.001**	0.18	3.71	**<0.001****
FA	-0.61	-16.60	**<0.001**	-0.21	-3.64	**<0.001****
NV	-0.30	-6.71	**<0.001**	-0.02	-0.61	0.545
PA	-0.49	-12.20	**<0.001**	-0.12	-2.73	**0.007****
DY	-0.33	-7.59	**<0.001**	-0.05	-1.23	0.221
SL	-0.38	-0.98	**<0.001**	-0.08	-1.91	0.057**
AP	-0.42	-9.88	**<0.001**	-0.05	-1.29	0.198
CO	-0.25	-5.64	**<0.001**	-0.04	-1.20	0.130
DI	-0.24	-5.43	**<0.001**	-0.09	-2.47	**0.014****
FI	-0.21	-4.52	**<0.001**	0.05	1.44	0.150

*; Standardized beta coefficient, **; selected by the backward elimination algorithm, ***; Dummy variables entered together
Quality of life dimensions: QL2; global quality of life, PF2; physical functioning, RF2; role functioning, EF; emotional functioning,
CF; cognitive functioning, SF; social functioning, FA; fatigue, NV; nausea and vomiting, PA; pain, DY; dyspnea, SL; insomnia,
AP; appetite loss, CO; constipation, DI; diarrhea, FI; financial difficulties.
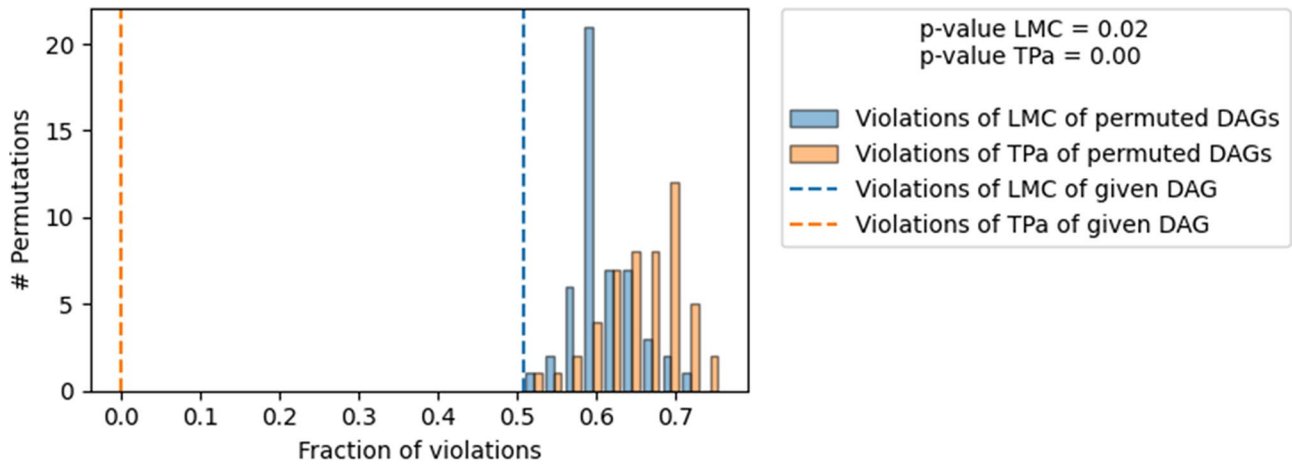
Fig. 2 Evaluation of directed acyclic graph (DAG). Directed acyclic graph structure. Directed acyclic graph (DAG) network of causal relationships among variables. Quality of life dimensions: NV, nausea and vomiting; SL, insomnia; AP, appetite loss; CO, constipation; DI, diarrhea; FI, financial difficulties; PF2, physical functioning; FA, fatigue; PA, pain; DY, dyspnea; RF2, role functioning; EF, emotional functioning; CF, cognitive functioning; QL2, global quality of life; SF, social functioning. Time_after_dx; time elapsed after diagnosis, diagx2_2d; type of cancer diagnosis (lung or colorectal cancer versus other cancers), active_treatment; whether the patient is on active treatment, stage; TNM stage of cancer

Fig. 3 Causal graph for global quality of life. Directed acyclic graph model evaluation. DAG: Directed acyclic graph. Two tests show that the DAG structure is explanatory

structure are not separately given for the purpose of brevity.

LiNGAM machine learning causal algorithm

With the Lingam algorithm, the adjacency matrix showed the network and direction of connections among the causal factors themselves and the causal factors directly and indirectly affecting QL2 and QL2 itself. There were 5 direct causal connections between various causal factors and QL2, namely, from EF to QL2, from FA to QL2, from PA to QL2, from RF2 to QL2, and from SF to QL2. The adjacency matrix is presented in Fig. 4. Interestingly, the causal effect direction in Fig. 4 between diagnosis and sex is from diagnosis to sex, whereas, we would expect the contrary, as difference in gender would be causing different cancers, such as female gender is expected to have more breast cancers, compared to the male gender. Here, apparently, the model wrongly guesses the direction of causal effect. In some cases, the relationship might be bidirectional or more complex than a simple cause-effect direction. LiNGAM might struggle with such relationships, leading to incorrect causal directions, which we think is the case here.

When total causal effects were considered, the most influential causal factor for QL2 was again the diagnosis (lung or colorectal cancer versus other cancers), as in the case with expert driven construction of the DAG. Figure 5 shows the causal effects of other factors on QL2. The main causal factors affecting QL2 revealed by manual DAG construction and the LiNGAM algorithm are detailed in Fig. 6, which shows the causal effects of diagnosis, stage, sex, nausea vomiting (NV), and SF on QL2.

Discussion

We have demonstrated, for the first time, that the type of cancer diagnosis, specifically the presence of lung or colorectal cancer versus other cancers, exhibits a causal relationship with global quality of life as shown by 2 separate methods of causal analysis. While some previous studies have indicated an association between cancer type and various quality of life scores, our analysis using machine learning and Bayesian inference techniques revealed that this relationship is not purely correlative but has a causal component [16]. In contrast, some authors have not shown a difference in global quality of life with respect to cancer type [17]. Interestingly,

Fig. 4 LiNGAM adjacency matrix of causal factors. Linear Non-Gaussian Acyclic Model (LiNGAM) adjacency matrix. Network of complex causal relationships among variables in the study. x0; age, x1; sex, x2; type of cancer diagnosis (lung or colorectal cancer versus other cancers), x3; TNM stage of cancer, x4; whether the patient is on active treatment, x5; time elapsed after diagnosis, x6; global quality of life, x7; physical functioning, x8; role functioning, x9; emotional functioning, x10; cognitive functioning, x11; social functioning, x12; fatigue, x13; nausea and vomiting, x14; pain, x15; dyspnea, x16; insomnia, x17; appetite, 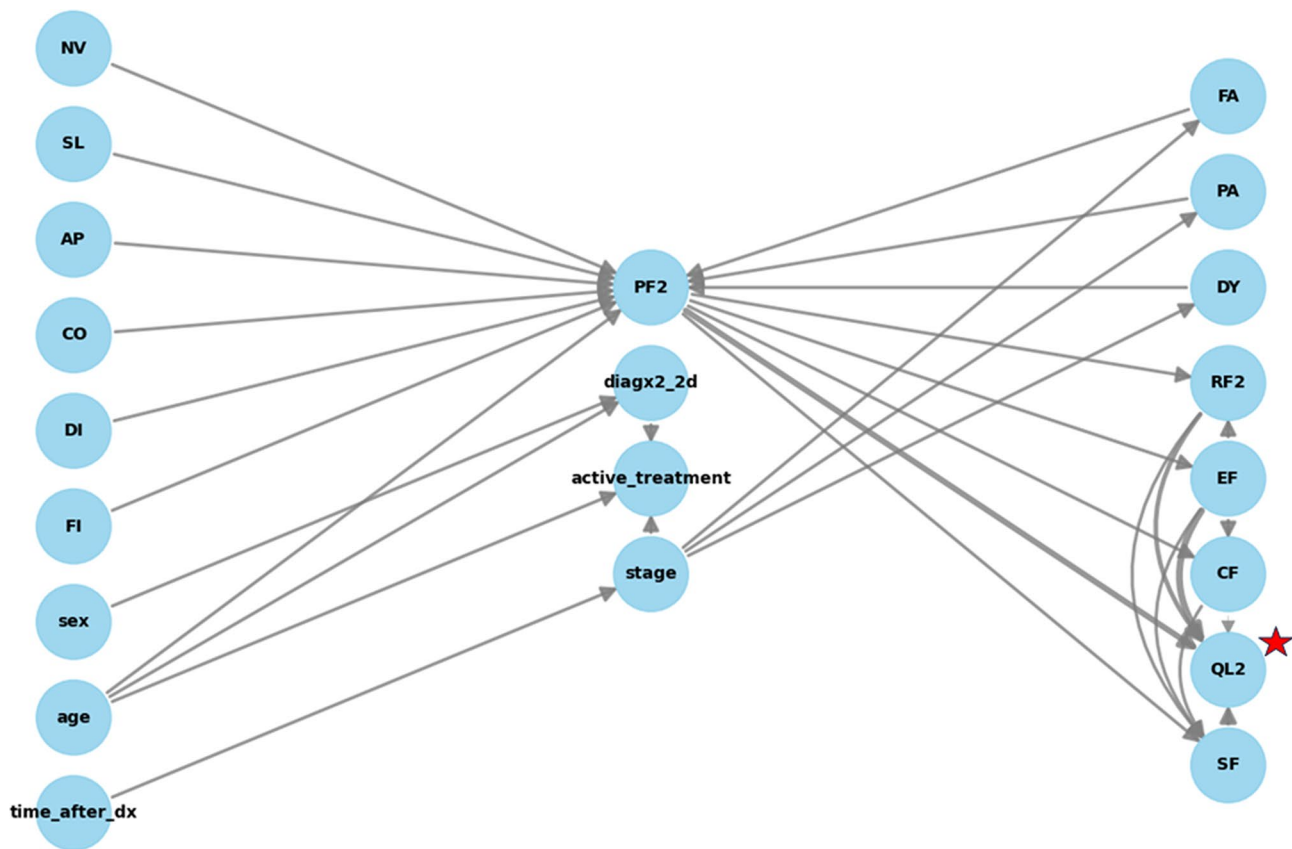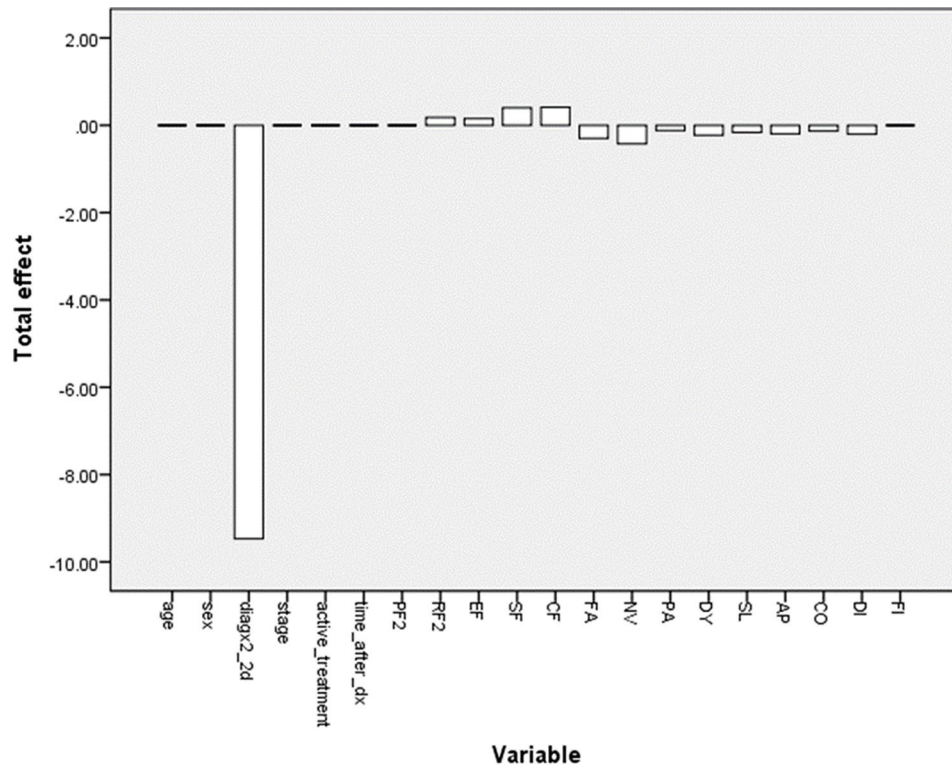x18; constipation, x19; diarrhea, x20; financial difficulties. Each node in the plot represents a variable, and the directed edges between nodes indicate the direction of causality. LiNGAM aims to identify the causal ordering of variables and their causal effects, so the plot will show the inferred causal structure based on the analysis of the data

Fig. 5 Total causal effects of LiNGAM. Linear Non-Gaussian Acyclic Model (LiNGAM) total causal effect scores. Abbreviations: diagx2_2d: lung or colorectal cancer, or other cancers, stage: cancer TNM stage at admission, active_treatment: receipt of chemotherapy or other systemic treatments, surgery or radiotherapy at the time of admission, time_after_dx: time elapsed since diagnosis at the time of admission, PF2: physical functioning, RF2: role functioning, EF: emotional functioning, SF: social functioning, CF: cognitive functioning, FA: fatigue, NV: nausea and vomiting, PA: pain, DY: dyspnea, SL: insomnia, AP: appetite loss, CO: constipation, Dl: diarrhea, FI: financial difficulties
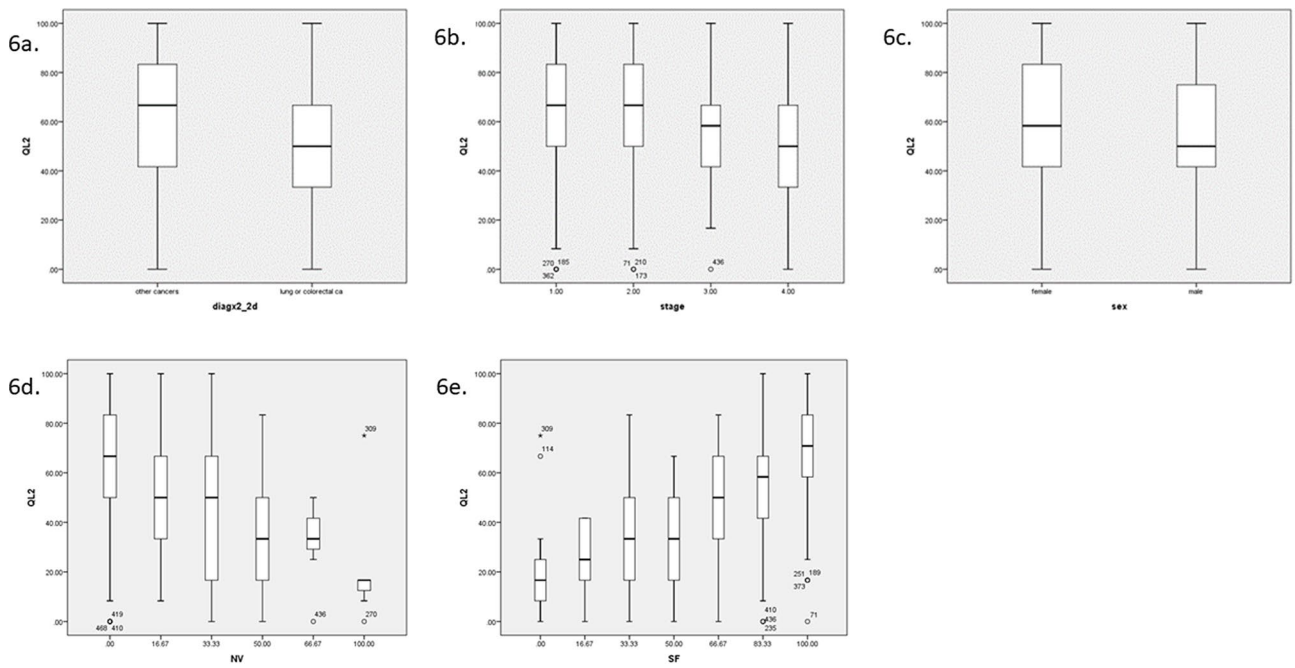
Fig. 6 Main total causal effects and global quality of life. Main total causal influences and global quality of life. Variables with the greatest total causal effects on global quality of life, as demonstrated in the Directed Acyclic Graph (DAG) and Linear Non-Gaussian Acyclic Models (LiNGAM). Global quality of life with reference to **6a**: diagx2_2d; type of cancer diagnosis (lung or colorectal cancer versus other cancers), **6b**: disease stage, **6c**: sex, **6d**: NV; nausea and vomiting, **6e**: SF; social functioning. Figure 6 includes diagnosis, stage and sex from DAG, and diagnosis, NV and SF from LiNGAM

the most influential causal factors for QL2 identified in this study differ from the strongest associates of QL2. Although the type of cancer diagnosis emerged as the most influential causal factor for QL2 in both the manual approach and the LiNGAM algorithm, it did not reach statistical significance in the multivariate analysis conducted using classical regression analysis. Similarly, while social functioning (SF) was the strongest associate of QL2 in the regression analysis, it did not emerge as the most influential causal factor in our causal analyses. This disparity underscores some unique insights that causal analysis could offer, revealing complex causal relationships that might not be captured by traditional associative or correlative approaches, such as regression analysis.

The DAG we manually coded did not include a visible path between the type of cancer diagnosis and QL2, but the total causal effect of the type of cancer diagnosis could be quantified algorithmically. However, looking at the LiNGAM adjacency matrix, we see that there are various possible paths between the type of cancer diagnosis and QL2, and thus, in that fashion, the LiNGAM adjacency matrix captured the complex relationships in our study better than the DAG. A possible explanation for the discrepancy between these 2 separate approaches (Expert-driven construction of the DAG versus LiNGAM) in our study is that the manually coded DAG failed to capture some of the important variables that were visible in the LiNGAM adjacency matrix. In other words, these variables remained latent for some of the relationships in the DAG structure, making LiNGAM, a machine learning model, more explanatory for our purpose.

The use of DAG and LiNGAM approaches in causal analysis offers several strengths and limitations. DAGs provide a clear and intuitive graphical representation of causal relationships, aiding in the identification of confounders, mediators, and potential sources of bias [18]. They facilitate the understanding of complex causal structures and the formulation of testable hypotheses. LiNGAM extends this by enabling the discovery of causal ordering and effects from observational data under the assumptions of linearity and non-Gaussian errors, offering a robust method for causal inference when these conditions are met. However, both approaches have limitations. DAGs rely heavily on domain knowledge for accurate construction, and their effectiveness is constrained by the correctness and completeness of this prior knowledge. LiNGAM's assumptions of linearity and non-Gaussianity may not always hold in real-world data, potentially limiting its applicability and accuracy. Additionally, LiNGAM can be sensitive to sample size and may struggle with high-dimensional data. Despite these limitations, the combined use of DAGs and LiNGAM provided a powerful toolkit for causal discovery and analysis for our study.

If the type of cancer directly influences global quality of life at the time of presentation, several factors may drive this phenomenon. Previous studies, including ours, have assessed quality of life across different cancer populations and have identified differences in various dimensions of quality of life; those cancer populations included metastatic or nonmetastatic cases, men or women, those with different categories of weight loss, performance status, weight loss, time from last treatment, and emotional and functional well-being [16, 19–21]. Specifically, the literature shows differences in the QL2 score between lung cancer patients and patients with other types of cancer [22]. Likewise, the quality-of-life score is also low in patients with colon cancer [23]. These differences are likely attributed to the distinct biological profiles of various cancers, with lung cancer often presenting at more advanced stages and having a poorer prognosis, especially in patients with stage 4 disease. Causal factors alter the magnitude of the affected variable, leading to a difference in its magnitude. This aligns with the basic concept of causality, where one variable (the cause) influences another variable (the effect). Thus, from this perspective, type of cancer causally effects and leads to difference in global quality of life in cancer patients. Additionally, the stigma associated with different types of cancer and perceptions of their curability may impact patients' experiences [24, 25]. For instance, a lung cancer patient may perceive a more unfavorable outlook than a patient with breast cancer. Thus, not only the biological and symptomatic aspects of the disease but also the patient's perception and societal view of the disease may influence the QL2 score. Further research focusing on the perception dimension and its relationship with quality-of-life dimensions across various cancers could provide valuable insights in this direction.

Apart from type of cancer, which has a causal effect on global quality of life, stage also emerged as another causal factor. Additionally, functional dimensions of QoL had positive causal effects, whereas symptom scales had negative causal effects on QoL2. Although there is some evidence that there is no association between stage or functional dimensions of QoL and QoL2, some available evidence suggests that some domains of QoL may be marginally different with respect to the disease stage [17, 21].

While our study provides important insights, it is not without limitations. Despite including data from 469 patients, a larger sample size, potentially in the range of thousands of cases, would be preferable for elucidating more intricate causal relationships and accurately assessing their strength. Additionally, obtaining more comprehensive data on social security, financial status, treatment details, and patients' perceptions of their disease status would enable a more thorough examination of the causal

effects of these factors in addition to the other variables considered in this study. Lastly, the validity of LiNGAM has been proven only for continuous variables; thus, our use of LiNGAM, which includes categorical or discrete data, is experimental.

The rise of AI and machine learning techniques is reshaping medical practice [26]. These technologies, including computer vision, natural language processing, and various machine learning algorithms, have been increasingly utilized for predicting outcomes and prognosis in recent years. Causal inference, whether through probabilistic, statistical, or machine learning-oriented algorithms, addresses the question of "Why did this happen?" rather than "What features or factors are associated with this outcome?" [27, 28]. This approach is expected to influence medical policies and interventions by providing deeper insights into causal relationships. Furthermore, causal inference techniques, such as causal survival analysis, are gaining traction in other research areas and are anticipated to play a more prominent role in medical research in the future [29]. The findings from our study lend support to this trend.

In summary, our study highlights the causal effect of cancer type, among other factors, on changes in QL2 scores in cancer patients. Importantly, this causal effect was not evident in classical regression analysis. The causal inference methodologies employed in our study have the potential to inform policies and interventions aimed at improving quality of life in cancer patients and, more broadly, to address numerous causal questions in oncology and medicine.

Conclusions

We found that cancer type is the primary causal factor for global quality of life in cancer patients. Additionally, our causal analyses revealed different factors compared to those identified in the regression analyses.

Abbreviations

LiNGAM	Linear Non-Gaussian Acyclic Model
QoL	Quality of life
EORTC QLQ-C30	European Organization for Research and Treatment of Cancer Quality of Life Questionnaire
AI	Artificial intelligence
QL2	global QoL
PF2	Physical functioning
RF2	Role functioning
EF	Emotional functioning
SF	Social functioning
CF	Cognitive functioning
FA	Fatigue
NV	Nausea and vomiting
PA	Pain
DY	Dyspnea
SL	Insomnia
AP	Appetite loss
CO	Constipation
DI	Diarrhea
FI	Financial difficulties
DAG	Directed acyclic graph
Time_after_dx	Time elapsed after diagnosis
Diagx2_2d	type of cancer diagnosis (lung or colorectal cancer versus other cancers)

Declarations

Ethics approval and consent to participate
The study was approved by the ethical committee of the Medical Park Hospital, Antalya, on the 17th of January 2024, with a reference number of 2024/1.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

References
1. Khan FA, Akhtar SS, Sheikh MK. Cancer treatment—objectives and quality of life issues. Malays J Med Sci. 2005;12(1):3–5.
2. Salas, et al. Health-related quality of life in women with breast cancer: a review of measures. BMC Cancer. 2022;22:66. https://doi.org/10.1186/s12885-021-09157-w.
3. Akhlaghi, et al. Health Qual Life Outcomes. 2020;18:33. https://doi.org/10.1186/s12955-020-01580-0.
4. Aaronson NK, Ahmedzai S, Bergman B, Bullinger M, Cull A, Duez NJ, Filiberti A, Flechtner H, Fleishman SB, de Haes JCJM, Kaasa S, Klee MC, Osoba D, Razavi D, Rofe PB, Schraub S, Sneeuw KCA, Sullivan M, Takeda F. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. J Natl Cancer Inst. 1993;85:365–76.
5. Kiebach J, Have JVT, Duijnhoven FJV, et al. Hematocrit is associated with cancer-related fatigue in colorectal cancer survivors: a longitudinal analysis. Cancer Epidemiol Biomarkers Prev. 2024 Jan;5. https://doi.org/10.1158/1055-9965.EPI-23-1048.
6. Hinz A, Mehnert A, Degi C, et al. The relationship between global and specific components of quality of life, assessed with the EORTC QLQ-C30 in a sample of 2019 cancer patients. Eur J Cancer Care. 2017;26(2):e12416.
7. Causal inference at https://en.wikipedia.org/wiki/Causal_inference. Accessed at 5.1.2024.
8. Beam AL, Drazen JM, Kohane IS, et al. Artificial Intelligence in Medicine. N Engl J Med. 2023;388:1220–1.
9. Zhang W, Ramezani R, Naeim A. Causal inference in medicine and in health policy, a summary. ArXiv, 2022: 2105.04655v4 [cs.LG].
10. Guzelant A, Goksel T, Ozkok S, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: an examination into the cultural validity and reliability of the Turkish version of the EORTC QLQ-C30. Eur J Cancer Care (Engl). 2004;13(2):135–44.

11. Sharma -Amit. Emre Kiciman. DoWhy: An End-to-End Library for Causal Inference. 2020. https://arxiv.org/abs/2011.04216.
12. Blöbaum P, Götz P, Budhathoki K, Mastakouri AA, Janzing D. DoWhy-GCM: an extension of DoWhy for causal inference in graphical causal models. 2022. https://arxiv.org/abs/2206.06821.
13. Ikeuchi T, Ide M, Zeng Y, Maeda TN, Shimizu S. Python package for causal discovery based on LiNGAM. J Mach Learn Res. 2023;24(14):1–8.
14. Shimizu S, Inazumi T, Sogawa Y, et al. DirectLiNGAM: a direct method for learning a linear non-gaussian structural equation model. J Mach Learn Res. 2011;12:1225–48.
15. Shimizu S, Hoyer PO, Hyvärinen A, Kerminen A. A linear non-gaussian acyclic model for causal discovery. J Mach Learn Res. 2006;7:2003–30.
16. Nguyen HTH, Duong KL, Nguyen ST, et al. Quality of life and its associated factors among cancer patients receiving chemotherapy at oncology hospitals in Vietnam after the third wave of the COVID-19 pandemic. Cancer Manage Res. 2022;14:2429–44.
17. Nayak MG, George A, Vidyasagar MS, et al. Quality of life among Cancer patients. Indian J Palliat Care. 2017;23(4):445–50.
18. Glymour C, Zhang K, Spirtes P. Review of causal discovery methods based on graphical models. Front Genet. 2019;10:1–15.
19. Daly LE, Dolan RD, Power DG, et al. Determinants of quality of life in patients with incurable cancer. Cancer. 2020;126(12):2872–82.
20. Han X, Robinson LA, Jensen RE, et al. Factors associated with Health-Related Quality of Life among cancer survivors in the United States. Cancer Spectr. 2021;5(1):pkaa123.
21. Ramasubbu SK, Pasricha RK, Nath UK, et al. Quality of life and factors affecting it in adult cancer patients undergoing cancer chemotherapy in a tertiary care hospital. Cancer Rep. 2020;4(2):e1312.
22. Polanski J, Polanska BJ, Rosinczuk J, et al. Quality of life of patients with lung cancer. Onco Targets Ther. 2016;9:1023–8.
23. Flyum IR, Mahic S, Grov EK, et al. Health-related quality of life in patients with colorectal cancer in the palliative phase: a systematic review and meta-analysis. BMC Palliat Care. 2021;20:144.
24. Yılmaz M, Dissiz G, Usluoglu AK, et al. Cancer-related stigma and depression in cancer patients in a middle-income country. Asia Pac J Oncol Nurs. 2020;7(1):95–102.
25. Knapp S, Marziliano A, Moyer A. Identity threat and stigma in cancer patients. Health Psychol Open. 2014;1(1):2055102914552281.
26. Rajpurkar P, Chen E, Banerjee O, et al. AI in health and medicine. Nat Med. 2022;28:31–8.
27. Ganguly N, Fazlija D, Badar M et al. A review of the role of causality in developing trustworthy AI systems. arXiv, 2023; 2302.06975v1 [cs.AI].
28. Sanchez P, Voisey JP, Xia T, et al. Causal machine learning for healthcare and precision medicine. R Soc Open Sci. 2022;9:220638. https://doi.org/10.1098/rsos.220638.
29. Murray EJ, Caniglia EC, Petito LC. Causal survival analysis: a guide to estimating intention-to-treat and per-protocol effects from randomized clinical trials with nonadherence. Res Methods Med Health Sci. 2021;2(1):39–49.

Publisher's Note