IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.0429000

Towards Interpretable Hybrid AI: Integrating Knowledge Graphs and Symbolic Reasoning in Medicine

YASHRAJSINH CHUDASAMA[2,3], HAO HUANG[2,3], DISHA PUROHIT[2,3], and MARIA-ESTHER VIDAL[1,2,3]

[1]L3S Research Center, Hannover, Germany
[2]Leibniz University of Hannover, Hannover, Germany
[3]TIB Leibniz Information Centre for Science and Technology, Hannover, Germany (e-mail:yashrajsinh.chudasama,hao.huang,disha.purohit,maria.vidal@tib.eu)

Corresponding author: Yashrajsinh (e-mail: yashrajsinh.chudasama@tib.eu), Disha (e-mail: disha.purohit@tib.eu), Hao (e-mail: hao.huang@tib.eu)

Yashrajsinh Chudasama, Hao Huang, and Disha Purohit equally contributed to this work and correspond to the first authors of this article. Yasharajsinh Chudasama, Hao Huang, Disha Purohit, and Maria-Esther Vidal are partially funded by the ''Leibniz Best Minds: Programme for Women Professors'', through funding of the ''TrustKG-Transforming Data in Trustable Insights'' project (Grant P99/2020).

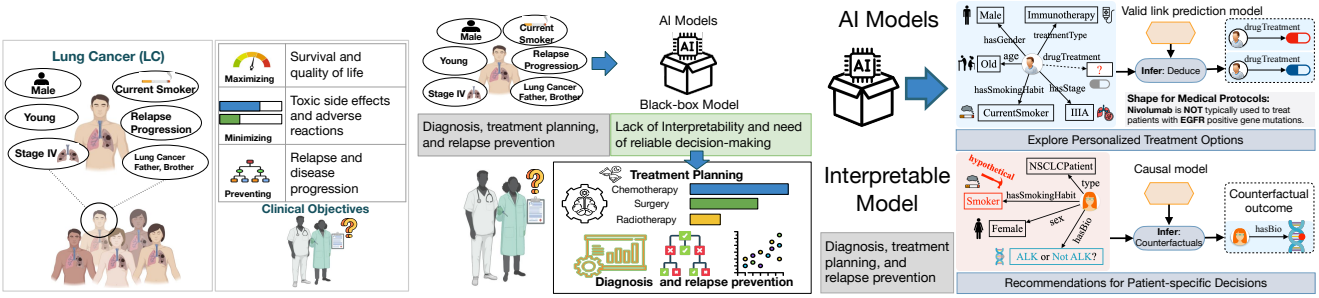
ABSTRACT Knowledge Graphs (KGs) are data structures that enable the integration of heterogeneous data sources and supporting both knowledge representation and formal reasoning. In this paper, we introduce TrustKG, a KG-based framework designed to enhance the interpretability and reliability of hybrid AI systems in healthcare. Positioned within the context of lung cancer, TrustKG supports link prediction, which uncovers hidden relationships within medical data, and counterfactual prediction, which explores alternative scenarios to understand causal factors. These tasks are addressed through two specialized hybrid AI systems, **VISE** and **HealthCareAI**, which combine symbolic reasoning with inductive learning over KGs to provide interpretable AI solutions for clinical decision-making. Leveraging KGs to represent biomedical properties and relationships, and augmenting them with learned patterns through symbolic reasoning, our hybrid approach produces models that are both accurate and transparent. This interpretability is particularly important in medical applications, where trust and reliability in AI-driven predictions are paramount.
Our empirical analysis demonstrates the effectiveness of **VISE** and **HealthCareAI** in improving the predictive accuracy and clarity of model outputs. By addressing challenges in link prediction—such as discovering previously unknown connections between medical entities—and in counterfactual prediction, TrustKG, with **VISE** and **HealthCareAI**, underscores the potential of integrating KGs with symbolic AI to create trustworthy, interpretable AI systems in healthcare. This paper contributes to the advancement of semantic AI, offering a pathway for robust and reliable AI solutions in clinical settings.

INDEX TERMS Counterfactual Prediction, Inductive Learning, Knowledge Graphs, Link Prediction, Symbolic Learning

I. INTRODUCTION

Knowledge graphs (KGs) provide a robust framework for integrating data and metadata from diverse sources [1]. By bridging disparate data silos and creating an integrated view, KGs enable powerful insights in complex domains that require comprehensive data analysis, such as healthcare. In lung cancer, for instance, diagnosis and treatment decisions are intricate, with oncologists needing to consider numerous factors—including tumor type, stage, and genomic markers [2]—to develop patient-specific treatment plans. These factors guide crucial decisions regarding therapies, dosages, and treatment cycles, which are tailored to individual patient responses and comorbidities [3], [4]. The primary clinical objectives in lung cancer management include: i) *Maximizing* survival and quality of life, ii) *Minimizing* side effects, and iii) *Preventing* relapse and disease progression [5].

While oncologists rely on their expertise and established guidelines, the variability in patient cases often makes it difficult to determine optimal treatment strategies through intuition alone. Predictive models could assist in areas such as diagnosis, treatment planning, and relapse prevention [2]–[4], [6]. However, traditional models frequently lack interpretabil-



(a) Clinical objectives include maximizing survival and quality of life, minimizing toxic side effects, and preventing relapse and disease progression.

(b) Limitations of traditional AI models in transparency and interpretability, requiring more reliable, understandable decision-making tools.

(c) Hybrid AI systems integrating symbolic reasoning and inductive learning over KGs provide interpretable and adaptable solutions, enabling personalized, actionable insights.

FIGURE 1: **AI Models in Medicine.** Figures (a) and (b) illustrate clinical objectives for LC treatment and the limitations in transparency of traditional AI models. Hybrid AI integrating symbolic reasoning with inductive learning over KGs presents a promising approach to overcoming these challenges, combining interpretability and adaptability to offer clinicians deeper tailored to specific patient needs (Figure (c)).

ity and domain-specific insights, which are crucial in lung cancer management where precision and transparency are essential. Figures 1a and 1b illustrate the clinical objectives for lung cancer treatment and the limitations in the transparency of traditional AI models in the healthcare domain.

To address these challenges, leveraging medical semantics can significantly enhance both the accuracy and interpretability of predictive models [7]. In particular, understanding *relapse risks* and *exploring personalized treatments* are essential for effective patient care in lung cancer [8]. These tasks can be approached through:

- *Relapse Risk Prediction through Valid Link Prediction*: Identifying latent connections between patient features (e.g., biomarkers) and relapse events helps to anticipate risk factors that may not be obvious in traditional datasets. This supports oncologists in preemptively adjusting treatment plans to mitigate relapse risks.
- *Personalized Treatment through Counterfactual Prediction*: Counterfactual reasoning allows the exploration of alternative treatment scenarios by simulating potential outcomes. This "what-if" analysis aids in identifying suitable treatment adjustments for individual patients, balancing relapse prevention and quality of life.

KGs capture complex clinical relationships, enabling valuable insights into tasks such as *relapse prediction* and *counterfactual prediction*, thereby contributing to improved patient care. Hybrid AI systems that integrate symbolic reasoning with inductive learning over KGs offer a promising approach to these challenges. By combining the interpretability of symbolic methods with the adaptability of inductive learning, hybrid AI can deliver deeper, more actionable insights tailored to specific patient needs (Figure 1c). This paper presents a methodology for integrating knowledge graphs (KGs) with inductive learning, which facilitates the development of interpretable, context-aware artificial intelligence (AI) in the healthcare domain.

A. OUR PROPOSED SOLUTION: TRUSTKG

We propose TrustKG to address these clinical tasks and deliver reliable, interpretable results to clinical users. TrustKG is a KG-based framework that supports semantic data management and KG analytics to promote interpretability in AI-driven healthcare. TrustKG facilitates tasks across data ingestion, processing, and integration, incorporating semantic alignment, named entity recognition, and entity linking/disambiguation. It includes tools for KG creation, validation, and exploration, enabling the representation of complex biomedical data with ontologies and logical reasoning, as well as validation through shape schemas. TrustKG's analytics capabilities leverage both symbolic and numerical learning techniques for predictive modeling, including valid link prediction, counterfactual prediction, and causal analysis.

Specifically, the hybrid AI systems **VISE** and **HealthCareAI** are integrated within TrustKG, combining symbolic reasoning with inductive learning over KGs to address the problems of valid link prediction and counterfactual prediction, respectively. **VISE** uncovers latent relationships in the data to assess relapse risks, while **HealthCareAI** uses counterfactual reasoning to explore personalized treatment options. Together, these systems enhance transparency and reliability in clinical decision-making by integrating interpretability with predictive precision. TrustKG also emphasizes human-centered communication by providing a natural language interface and referencing scientific publications to support its recommendations, empowering clinicians to make informed, evidence-based decisions.

B. CONTRIBUTIONS AND STRUCTURE

Building on our previous work on **VISE** [9] and **HealthCareAI** [10], this paper makes the following novel contributions:

- The development of TrustKG, an interoperable and interpretable knowledge graph (KG) ecosystem tailored for

the medical domain.
- The problems of *Valid Link Prediction* and *Counterfactual Prediction*.
- The specification of **VISE** and **HealthCareAI**, as hybrid AI systems, which address these problems using symbolic reasoning combined with inductive learning to improve prediction accuracy in lung cancer.
- Empirical evaluation of **VISE** and **HealthCareAI** on KGs created from a lung cancer dataset, demonstrates the impact of semantics on predictive performance.

The paper is structured as follows: Section II reviews related work, while Section III describes the main features implemented in TrustKG. Section IV defines the problem of valid link prediction and presents **VISE**, along with the results of an experimental evaluation of its performance. Similarly, Section V introduces the problem of counterfactual prediction and positions **HealthCareAI** as a hybrid AI system that leverages semantics to effectively solve this task. Finally, Section VI provides concluding insights and future directions.

II. RELATED WORK

A. KGS IN MEDICINE

In the field of medicine, Knowledge Graphs (KGs) have been widely employed to address challenges in semantic data integration through approaches like knowledge extraction [11] and exploration [12]. Chandak et al. [13] introduce PrimeKG, a KG that unifies clinical concepts from diverse medical sources, encompassing 17,080 diseases with alignments to biological processes, experimental drugs, and protein alterations. PrimeKG enables forecasting of drug-disease interactions and treatment recommendations. Sakor et al. [14] present Knowledge4COVID-19, a framework for constructing healthcare KGs using rule-based entity linking and machine learning. This KG facilitates tasks such as drug-drug interaction prediction, treatment recommendations for COVID-19, and treatment impact visualization. Knowledge4COVID-19 showcases the potential of structured clinical knowledge extraction from unstructured data. In oncology, Fotis et al. [11] propose DE4LungCancer, a health data ecosystem leveraging controlled vocabularies and ontologies to represent lung cancer patients' medical histories. The ecosystem uses the RML mapping engine [15] for KG creation from diverse data sources, and SHACL technologies for data quality validation. Trav-SHACL [16], an efficient SHACL validation engine, assesses entities (e.g., patients) against specific medical protocols. Ristoski et al. [17] explore KG-based data linking and integration, supporting effective data discovery and knowledge exploration. Rivas et al. [18] propose an inductive learning approach using graph neural networks to detect drug-drug and drug-target interactions, predicting molecular properties and potential treatments. Callahan et al. [19] introduce an open-source KG ecosystem for life sciences, presenting a methodology to construct large-scale KGs with components for data pre-processing, KG construction, and analytics. In lung cancer research, Calvo et al. [20] develop the TTR KG, which integrates knowledge on non-small cell lung cancer, including data from 12,351 patients with attributes such as smoking habits and mortality rates. The TTR KG provides analytical insights and supports survival rate predictions. Empirical evaluation of **VISE** and **HealthCareAI** on KGs created from a lung cancer dataset, demonstrating the impact of semantic integration on predictive performance.

B. AI MODELS IN MEDICINE

Recent advancements in AI have brought personalized healthcare systems closer to reality. Janik et al. [21] propose a personalized care model for lung cancer patients to estimate relapse probability, comparing the accuracy of traditional AI models (78%) with graph machine learning models (68%) for relapse prediction. Pan et al. [22] investigate the effectiveness of AI models, such as Support Vector Machine and Random Forest, in predicting relapse for acute lymphoblastic leukemia patients. Similarly, Yang et al. [23] explore how machine learning techniques, including decision trees and deep neural networks, can analyze the influence of clinical status and demographics on the survivability of early-stage cancer patients. Additional studies [24], [25] examine AI techniques for predicting lung cancer mutations from tabular and image data, highlighting the impact of attributes like prescribed treatments on predictive accuracy. One study closely related to ours explores the use of machine learning models, particularly random forests [26], in personalized healthcare. Vyas et al. [27] model patient-level and patient-episode health records, developing ensemble-based predictive models for dementia prognosis and personalized treatment recommendations, with interpretability provided by LIME [28]. This approach helps oncologists forecast and recommend treatments, enhancing decision-making in clinical settings. While these AI models show promise for predictive tasks in medicine and oncology, they often lack interpretability and do not fully leverage domain-specific knowledge, which limits their reliability in clinical settings. Integrating semantic information through KGs offers a pathway to address these limitations, enhancing interpretability and enabling AI systems to provide more context-aware insights for personalized patient care.

C. EXPLAINABLE AI IN MEDICINE

As AI adoption in medicine grows, a key challenge persists: many models function as black boxes, making their decision processes opaque and difficult to interpret. Understanding the rationale behind AI predictions is essential for trust and reliability, especially in healthcare. Suh et al. [29] highlight the communication gap between AI practitioners and medical experts, showing that accuracy metrics alone are insufficient for building understanding among subject matter experts. Instead, models must be explicitly interpretable to provide meaningful insights. To address this need, Li et al. [30] identify distinct user personas in Knowledge Graph (KG) applications—KG Builder, KG Analyst, and KG Consumer—each with unique requirements for interpretability. Similarly, Purohit et al. [31] propose the *DIGGER* pipeline, which extracts logical rules from lung cancer treatment data to

reveal patterns, flag protocol deviations, and complete missing relationships within KGs, providing clinicians with transparent, explainable insights. Neuro-symbolic AI approaches have furthered explainability by combining symbolic and numerical methods. For instance, Rivas et al. [32] predict therapy efficacy using a hybrid learning approach, while Chudasama et al. [33], [34] introduce InterpretME, a KG-based framework that traces and explains predictive model outcomes with LIME and SHAP [35]. In the SemDesLC framework [36], Semantic Web technologies are leveraged to make lung cancer relapse predictions interpretable, bridging gaps between medical insights and AI predictions. Despite these advancements, limitations remain. Current models often lack domain-specific semantics, restricting their ability to capture the complex medical context needed for clinical decision-making. Many systems also struggle to integrate heterogeneous data effectively, limiting the interpretability of complex relationships. These challenges highlight the need for frameworks like TrustKG, which integrates symbolic reasoning with inductive learning over KGs, enabling semantically enriched, context-aware insights that support clinical decisions in complex healthcare scenarios.

D. AI MODELS FOR CAUSAL REASONING

Advanced predictive models have been widely used in medicine for diagnosis support, prognosis, and personalized treatment recommendations [37]–[39]. However, while these models provide valuable predictions, they often lack interpretability and trustworthiness—key elements in clinical decision-making where understanding the "why" and "what if" behind recommendations is crucial [40], [41]. Causal models, which address these questions, require assumptions about dependencies among clinical variables, such as the Stable Unit Treatment Value Assumption (SUTVA) and the Ignorability Assumption [42]. To enhance causal analysis, Salimi et al. [43] developed CaRL, a framework for causal reasoning over relational databases, while Huang et al. proposed CareKG and CauseKG to model causal relationships and estimate causal effects in KGs [44], [45]. However, these models fall short in handling counterfactual prediction ("what if" scenarios), which requires comprehensive causal graphs (CGs) to capture all relationships. Techniques like structural causal models (SCMs) and causal Bayesian networks (CBNs) allow counterfactual predictions when a CG is available, but CG construction is challenging and often relies on domain experts or data-driven methods like the Peter-Clark (PC) algorithm and Greedy Equivalence Search (GES) [46]. Despite recent advances, generating complete causal graphs from data remains theoretically unattainable [40]. Emerging methods using large language models for causal discovery still lack integration of clinical semantics, which could improve CG accuracy. Additionally, Barbra et al. [47] explore KG embeddings for counterfactual prediction without a CG, but their real-world effectiveness is yet to be established. Moreover, existing causal models in healthcare lack comprehensive semantic integration, limiting their interpretability. We address these gaps by integrating semantic reasoning with KGs, enabling richer causal insights and reliable counterfactual predictions tailored for complex clinical decision-making.

III. KGS AND AI MODELS IN MEDICINE

The application of AI in medicine, particularly in complex fields like oncology, requires systems that not only provide accurate predictions but also meet high standards of interpretability, reliability, and personalization. KGs offer a powerful approach for integrating heterogeneous medical data and encoding biomedical knowledge, making them ideal for use in hybrid AI systems. This section discusses the essential requirements for AI in medical applications, the structure and capabilities of TrustKG—a hybrid AI framework based on KGs—and introduces two key predictive tasks, valid link prediction and counterfactual prediction, relevant to improving decision-making in lung cancer treatment.

A. REQUIREMENTS FOR AI IN MEDICINE

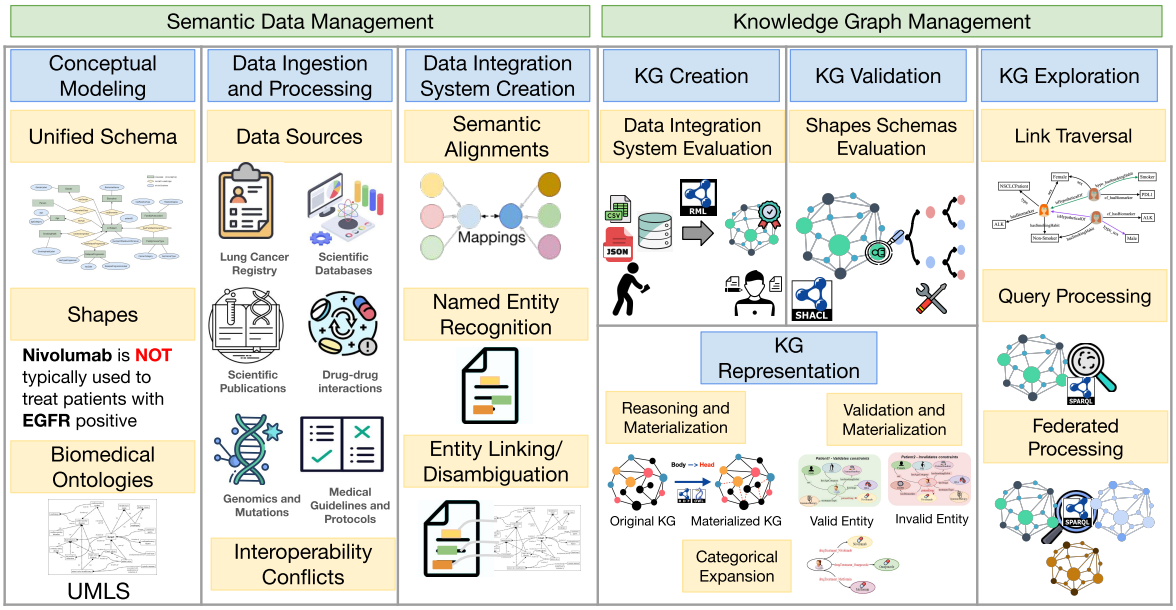
The implementation of effective AI models to address predictive problems in medicine requires the satisfaction of several key requirements [3], [36], [48]:

- *Interpretability*: Oncologists need to understand the rationale behind AI-driven recommendations. This is crucial for validating model suggestions and making adjustments based on individual patient needs [49].
- *Reliability*: An AI model should provide consistent, accurate predictions, reducing the risk of erroneous recommendations that could negatively impact patient outcomes [50].
- *Context-Aware Reasoning*: AI models must incorporate medical knowledge, such as relationships between patient characteristics, tumor markers, and treatment outcomes. This knowledge enables AI systems to make decisions that align with clinical guidelines and patient-specific factors [51].
- *Personalization*: Effective AI models must be able to personalize treatment recommendations and account for patient-specific factors, such as genetic mutations, medical history, and potential comorbidities [3].
- *Adaptability*: AI systems should adapt to new medical knowledge and evolving treatment protocols. This requires a flexible framework capable of integrating updated data and knowledge sources over time [52].

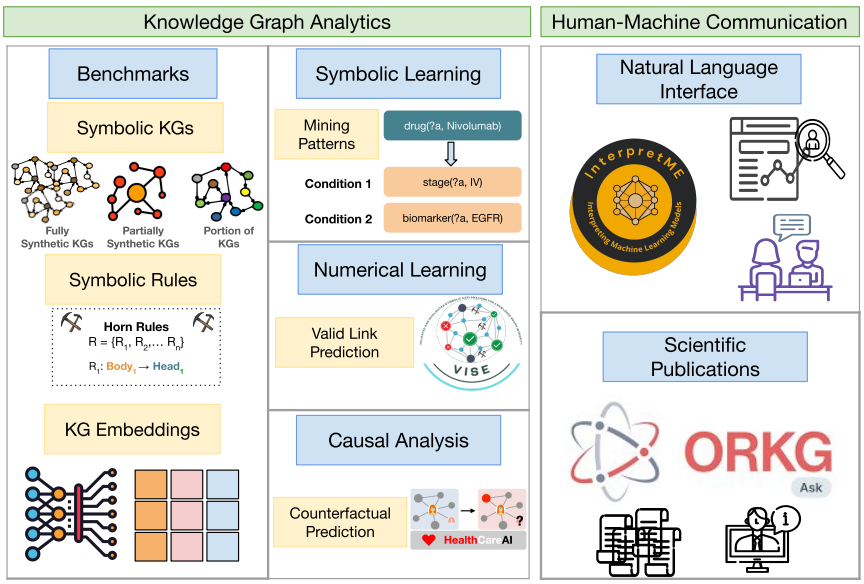
Meeting these requirements is challenging with traditional machine learning approaches, as they often lack the necessary transparency and domain-specific knowledge [53]. To address these limitations, by combing semantics inferred from symbolic reasoning and integrity constraint validation with inductive learning over KGs.

B. TRUSTKG: A HYBRID AI FRAMEWORK FOR MEDICAL KNOWLEDGE GRAPHS

In the medical domain, the complexity of patient data requires AI systems that can integrate diverse, domain-specific knowledge while maintaining interpretability, reliability, and

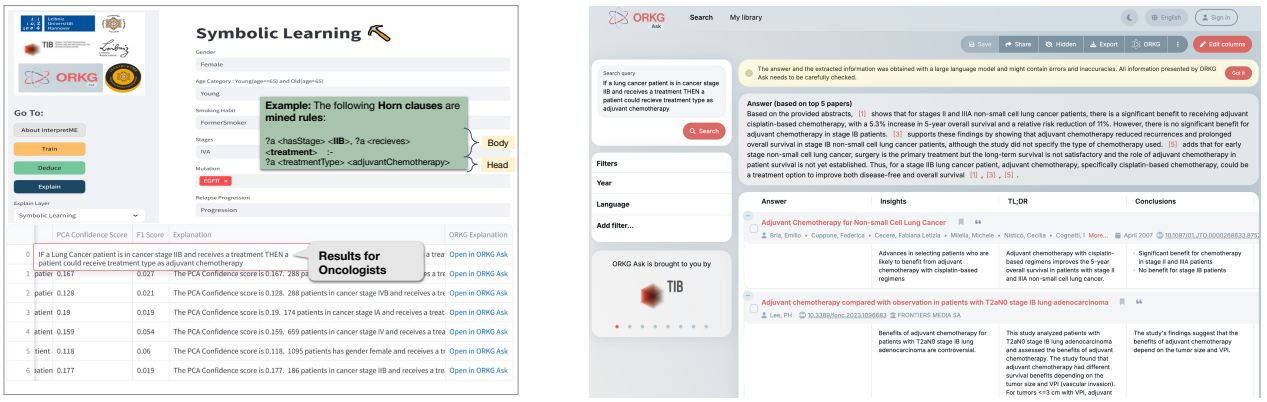


(a) **Semantic Data Management and Knowledge Graph Management in TrustKG.** The Semantic Data Management layer supports conceptual modeling, data ingestion, processing, and integration across heterogeneous data sources, utilizing techniques like semantic alignment, named entity recognition, and entity linking to harmonize medical data. The Knowledge Graph Management layer focuses on creating, validating, and exploring KGs, with tools for reasoning, shape validation, and federated processing to ensure the integrity and usability of the graph structure. These capabilities are essential for structuring complex biomedical knowledge accurately, providing a foundation for reliable AI-driven insights in clinical applications.



(b) **Knowledge Graph Analytics and Human-Machine Communication in TrustKG.** The Knowledge Graph Analytics layer combines symbolic learning, numerical learning, and causal analysis, enabling tasks such as valid link prediction and counterfactual prediction. The Human-Machine Communication layer provides a natural language interface and integration with scientific publications, making it easier for clinical users to interpret AI-driven recommendations. Thus, enabling TrustKG to offer interpretable and actionable insights, bridging the gap between complex data analytics and real-world clinical decision-making, and demonstrating the practical value of semantic AI in healthcare.

FIGURE 2: *TrustKG Framework:* This architecture allows TrustKG to deliver accurate, interpretable, and clinically relevant insights, supporting evidence-based decision-making in complex medical settings.



(a) InterpretME Interface

(b) ORKG Ask Interface

FIGURE 3: **Human-Machine Communication in TrustKG**. **(a) InterpretME Interface:** The interface displays mined Horn rules (e.g., treatment recommendations based on cancer stage and mutation type) with their corresponding confidence scores and explanations. These results are transformed into natural language, making them accessible to oncologists for interpretation and decision-making. Additionally, links to ORKG explanations provide further insights into the derived rules. **(b) ORKG Ask Interface:** The ORKG Ask system retrieves relevant scientific publications and presents query results (e.g., chemotherapy treatment outcomes) along with concise summaries, insights, and links to full articles. This enables oncologists to access evidence-based findings directly connected to the mined rules, supporting informed decision-making.

adaptability. Knowledge Graphs (KGs) are particularly suited to this challenge. A KG is a directed, edge-labeled graph that can be represented as $KG = (V, E, L)$ [54], where V represents nodes (e.g., entities like patients, treatments, or biomarkers), L denotes edge labels (e.g., relationships such as "has_symptom" or "treated_with"), and E is the set of edges connecting nodes based on these relationships. KGs enable the representation of medical knowledge, capturing complex interrelationships essential for accurate and context-aware AI-driven insights. To meet the requirements of interpretability, reliability, context-aware reasoning, personalization, and adaptability in clinical applications, we propose **TrustKG**, a hybrid AI framework that leverages KGs to support advanced predictive tasks in healthcare. TrustKG integrates symbolic reasoning with inductive learning, to provide clinically relevant insights. As illustrated in TrustKG is organized into four core layers—Semantic Data Management, Knowledge Graph Management, Knowledge Graph Analytics, and Human-Machine Communication (Figures 2a and 2b)—that together address challenges of medical data integration, knowledge management, interpretation, and usability.

Semantic Data Management facilitates the ingestion, processing, and integration of diverse data sources, including scientific literature, clinical guidelines, and patient records. This layer achieves data harmonization through semantic alignment, named entity recognition, and entity linking, creating a unified and interoperable schema that captures complex biomedical knowledge within an integrated data system. This structured approach enables the mapping of data sources to unified schemas or biomedical ontologies (e.g., the Unified Medical Language System - UMLS[1]), while clinical protocols and guidelines are encoded as shape constraints. It is of paramount importance that domain-specific representations are employed to guarantee that the outputs produced by TrustKG are both interpretable and contextually relevant.

Knowledge Graph Management supports the creation, validation, and exploration of KGs, enabling reasoning and materialization. It uses shape schemas to validate the KG, ensuring the consistency and reliability of represented entities and relationships, and offering clinicians trustworthy insights. Federated query processing allows seamless exploration across multiple datasets, enhancing the system's adaptability and usability in healthcare environments.

Knowledge Graph Analytics combines symbolic and numerical learning to enable predictive modeling. Symbolic learning identifies patterns within the graph, while numerical learning supports tasks such as valid link prediction and causal analysis. These capabilities directly address the need for context-aware reasoning by drawing insights from established medical knowledge and real-world data. In lung cancer treatment, for example, these analytics support tasks such as assessing risk of relapse and exploring treatment scenarios. In addition, various techniques will be implemented to generate benchmarks of synthetic KGs (i.e., fully and partially synthetic KGs [55]) for the empirical studies of TrustKG tools.

Human-Machine Communication provides a natural language interface and links insights to scientific publications, enhancing interpretability and trust. This layer ensures that clinicians receive recommendations in accessible language, allowing them to validate insights against existing literature

[1]Accessed on November 14th, 2024: https://www.nlm.nih.gov/research/umls/quickstart.html

and make informed decisions. By presenting results transparently, TrustKG fosters clinician confidence in the AI's output. Additionally, TrustKG leverages frameworks such as InterpretME [33], a KG-driven system that enables fine-grained representation of the main characteristics of trained machine learning models, ensuring that clinicians and researchers can better understand model behaviors and predictions. TrustKG also integrates tools like ORKG Ask [56], a hybrid AI system that supported on vector search, large language models, and KGs allows for the scholarly search and exploration. In TrustKG, ORKG Ask provides advanced query and exploration capabilities, facilitating access to relevant scientific knowledge and enhancing clinicians' ability to interpret AI-driven insights in a broader research landscape.

C. PREDICTION PROBLEMS IN TRUSTKG

TrustKG addresses two core prediction tasks critical in oncology: *Valid Link Prediction* and *Counterfactual Prediction*. These tasks are of great importance for the implementation of personalized treatment strategies and the anticipation of patient outcomes in complex scenarios.

Valid Link Prediction: This task involves assessing the likelihood that a given KG link is accurate. In lung cancer, this can mean predicting the probability of relapse based on biomarkers and other patient-specific factors. Accurate link prediction enables early intervention and improved prognosis by identifying at-risk patients. In the context of lung cancer, *Valid Link Prediction* is illustrated by assessing whether a prescribed treatment, such as immunotherapy, is valid for a specific patient profile. The example in Figure 4a shows a patient with attributes including age, gender, smoking habits, and cancer stage. The predictive model evaluates if the patient's drug treatment aligns with medical protocols, such as whether Nivolumab is recommended for patients without EGFR-positive mutations. By anticipating the efficacy of this treatment link, TrustKG facilitates the delivery of optimal care, promoting early intervention and enhanced outcomes.

Counterfactual Prediction: Counterfactual reasoning predicts outcomes under hypothetical scenarios, providing a "what-if" analysis. In lung cancer, counterfactual predictions help evaluate how changes in lifestyle factors, like smoking cessation, might influence disease progression. This capability is essential for personalized treatment planning and preventive care. Counterfactual Prediction enables "what-if" analyses by predicting outcomes under hypothetical scenarios. In the example shown in Figure 4b, the model examines how changes in lifestyle factors, such as smoking habits, might influence a biomarker status related to lung cancer. For instance, it assesses if switching from a current smoker to a non-smoker could affect the ALK mutation status. This analysis supports clinicians in evaluating the potential benefits of lifestyle changes, aiding more in personalized treatment planning, emergency response planning, and preventive care by providing a deeper understanding of patient-specific risks.

TrustKG integrates two specialized hybrid AI systems within: **VISE** for valid link prediction and **HealthCareAI** for counterfactual prediction. Using design patterns and structured implementation frameworks, both systems combine symbolic reasoning with inductive learning over KGs. **VISE** supports relapse risk assessment by uncovering latent connections, while **HealthCareAI** enables practitioners to explore alternative treatment scenarios tailored to individual needs. Together, **VISE** and **HealthCareAI** illustrate the potential of TrustKG to improve decision-making in healthcare by providing interpretable and personalized insights that support both immediate and preventive medical actions.

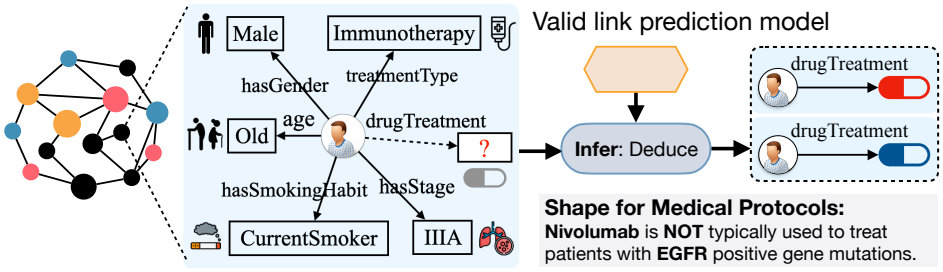
IV. PREDICTING VALID LINKS OVER KGS

This section defines the problem of valid link prediction and presents **VISE**. The boxology of design patterns proposed by Van Bekkum et al. [57] is utilized to specify the main components of **VISE**. Further, the impact of considering semantics encoded in integrity constraints is empirically evaluated, and the experimental results are reported. We start summarizing basic concepts like *shapes*, *shapes schemas*, *inductive learning*, *Partial Completeness Assumption* heuristic and score, and *Design Patterns*.

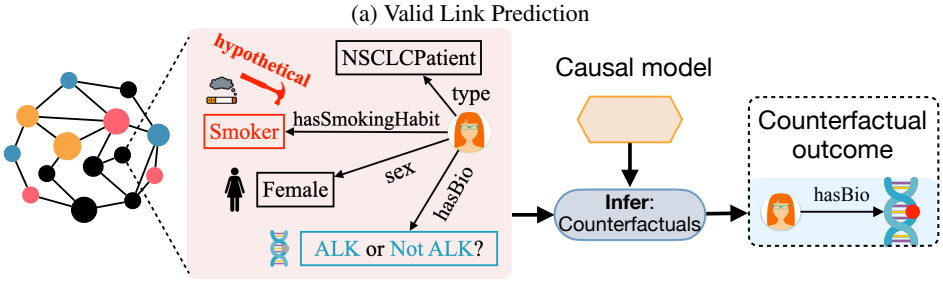
Shapes and Shapes Schemas. *Shapes* represent sets of conditions that nodes or edges must satisfy to ensure data consistency and integrity in KGs. A shape, denoted as ϕ, can include basic conditions like truth values, membership tests, Boolean conditions, conjunctions, negations, or cardinality constraints on edges. *Shapes schemas*, denoted as $\Sigma = (\varphi, S, \lambda)$, organize these shapes into a framework consisting of a set of shapes (φ), labels (S) for identification, and a mapping function (λ) linking labels to shapes. Evaluation of a shape determines whether a node or edge satisfies the defined constraints, yielding a binary result (0 or 1). A KG satisfies a shapes schema if all nodes targeted by the schema validate their respective shapes. This structured approach provides a foundation for encoding and enforcing constraints in KGs.

Inductive Learning Over KGs. Inductive learning in Knowledge Graphs (KGs) derives general patterns from specific data to make predictions about unseen facts [54]. It employs methods such as *Knowledge Graph Embeddings*, which map KGs to low-dimensional vector spaces to capture relationships numerically, and *Symbolic Learning*, which identifies logical rules (e.g., Horn clauses) denoted as R for symbolic reasoning. To address the incompleteness nature of KGs, the *Partial Completeness Assumption (PCA)* heuristic integrates observed positive edges (E^+) with *heuristic-based negative edges (hE^-)*. These hE^- edges represent plausible but unobserved relationships entailed from the logical structure of rules. The *PCA Confidence Score*, $PCA(R) = \frac{\text{support}(R)}{|E^+ \cup hE^-|}$ measures a rule's reliability by balancing its support based on edges entailed from R, against the union of observed (E^+) and entailed edges (hE^-).

Boxology of Design Patterns. Van Bekkum et al. [57] introduced modular design patterns, known as "boxology," to describe hybrid neuro-symbolic systems using elementary design patterns. These patterns visually represent system components, with inputs/outputs as grey rectangles, processes



(a) Valid Link Prediction



(b) Counterfactual Prediction

FIGURE 4: **Prediction Tasks in TrustKG.** (a) Valid Link Prediction: An example illustrating the identification of a link between patient attributes and treatment protocols. The model deduces whether the prescribed treatment aligns with clinical guidelines, indicating if it is valid for a patient with specific characteristics (e.g., age, gender, and genetic mutations). (b) Counterfactual Prediction: An example of counterfactual reasoning applied to evaluate how a hypothetical scenario (e.g., change in smoking habits) might impact the patient's biomarker profile, supporting personalized treatment decisions and preventive care.

as blue rounded rectangles, models as yellow hexagons, and actors as green triangles. Key patterns for generating models include: *Training*, where data or symbols are used to create a model; *Transformation*, where data or symbols are converted into new forms without creating a model; and *Inference for Symbols*, where a model deduces symbols from data or symbols. They are applied to the proposed solutions for predicting valid links and counterfactual prediction (section V).

A. PROBLEM STATEMENT

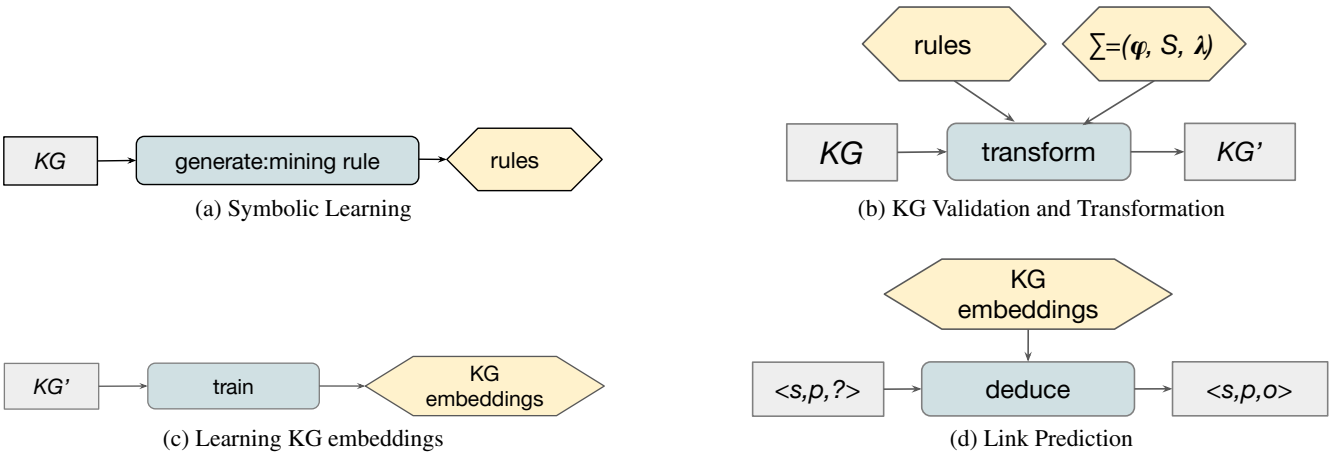
Given a directed edge-labeled graph *KG=(V, E, L)* where each node $v \in V$ represents an entity and each $p \in L$ represents a unique relation between entities, the task of link prediction focuses on completing an incomplete triple $\langle s, p, ? \rangle$. Specifically, this involves identifying the most plausible entity $o' \in V$ that completes the triple (s, p, o), such that both s and o' validate a given shape schema $\Sigma = (\varphi, S, \lambda)$. Using a scoring function $\theta(s, p, o')$, which quantifies the plausibility of the triple, the optimization problem is defined as:

$$o' = \arg\min_{e \in V} \theta(s,\ p,\ e) \wedge s \models \varphi \wedge e \models \varphi$$

The goal is to infer the most plausible entities o' by leveraging positive edges E^+, generating heuristic-based negative edges hE^-, and ensuring that the resulting triples satisfy the shape schema Σ, i.e., $(s, p, o') \models \varphi$.

B. PROPOSED SOLUTION

We propose **VISE** to solve the problem of *Valid Link Prediction*. **VISE** follows the hybrid design pattern as illustrated in Figure 5, strategically combining numerical learning with symbolic learning and constraints validation methods. Figure 5a illustrates *Symbolic learning* component which is applied to the input KG, resulting in the generation of logical rules and PCA heuristic-based edges. For instance, a logical rule, *relapseProgression(?a, Progression), drug(?a, Nivolumab)* $\Rightarrow$ *biomarker(?a, EGFR_Negative)* stating that if a patient has prescribed nivolumab in the progression is more likely to have EGFR Negative mutation. The learned heuristic-based edges serve as prior knowledge, improving numerical learning approaches such as KGE models combined with constraints validation and KG transformation. During the process of symbolic learning, **VISE** utilizes extracted Horn rules in conjunction with PCA Confidence to deduce heuristic-based negative edges. The mined rules are subsequently utilized to generate predictions regarding the missing relationships in the input KG. These predictions are based on logical inference, which is used to calculate the entailment of the mined rules. SPARQL queries are employed to infer the entailment of mined horn rules and generate heuristic-based negative edges (hE^-) which represent the deduced knowledge based on the observed explicit patterns. The predictions generated by the symbolic learning system in conjunction with the input KG are then fed to the *KG Validation and Transformation* (Figure 5b) component, where the predicted links are evaluated to determine whether they validate or invalidate the SHACL constraints. Here, the validation framework is defined by $\sum = (\varphi, S, \lambda)$, a symbolic component to ensure the quality of the predicted hE^- edges. For



(a) Symbolic Learning
(b) KG Validation and Transformation
(c) Learning KG embeddings
(d) Link Prediction

FIGURE 5: **VISE** (corresponds to the Numerical Learning component in Figure 2). Hybrid design pattern (Figure 5a to 5d), demonstrates the use of symbolic rules from the Symbolic Learning component and Constraint Validation and KG Transformation component in combination with Numerical Learning (i.e., KGE models) to enhance the predictive performance.

instance, a SHACL constraint is to check *if a patient mutated with EGFR is not recommended Afatinib drug*. Furthermore, the generated validation report is utilized to transform or rewrite the KG with categorical representation to contain the symbolic knowledge resulting from the constraints validation. The transformed KG is then provided as input to the numerical learning models, i.e., KG embedding models, during the training phase (Figure 5c). This is achieved by processing the data into a low-dimensional vector space. The process of numerical learning is capable of predicting missing links, thereby completing the KGs by deducing links that validate the constraints at a higher rank and with a greater probability of accuracy (Figure 5d). *Categorical representation of KGs* before giving as input to the numerical learning component transforms the KGs to contain negated facts, allowing the KGE model to learn all the representations, enhancing the performance of the models and empowering KG completion. Several studies demonstrated the need for negated facts in KGs to boost the performance of KGE models. **VISE** employs a two-fold rewriting process. Firstly, it evaluates the links predicted by symbolic learning using constraints. Second, depending upon the validation report of the predicted link. If the patient in the lung cancer KG invalidates the constraint, the links that resemble the patient characteristics in the KG are added with negation, i.e., negated facts. Furthermore, each component builds upon the previous one, creating a fully-fledged hybrid framework, **VISE**, that encompasses enrichment, validation, and transformation of deduced knowledge based on observed explicit patterns to infer meaningful predictions.

Figure 6 illustrates each of the **VISE** components. During *Symbolic Learning* (a), where mining rules are generated. In this case, a rule suggests that patients receiving *Nivolumab* for *stage IIIA* lung cancer as part of *immunotherapy* are more likely to exhibit certain biomarkers, such as *EGFR Negative*. In the *KG Validation and Transformation* step (b),

TABLE 1: **Benchmark Statistics** represents triple counts along with counts of entities, relations, and existing known-true-and-false facts.

Synthetic Lung Cancer (SLC) KG			
#triples	20581	**#known-true**	200
#entities	383	**#known-false**	170
#relations	43		

SHACL constraints are applied to validate predicted links. For instance, a constraint checks whether patients with an *EGFR* mutation are recommended against *Nivolumab*. The rules and constraints are used to transform the initial KG into an enriched KG' with both positive and heuristic-based negative edges. The *Learning KG Embeddings* step (c) trains a KG embedding model on the transformed KG', capturing both validated and negated relationships in a low-dimensional representation. Finally, in the *Link Prediction* phase (d), the embeddings are used to deduce new plausible links, such as predicting the likelihood of a patient having *relapse progression* based on existing patterns. This pipeline shows how **VISE** combines symbolic rules, constraints validation, and numerical learning to enhance prediction accuracy in KGs.

C. EXPERIMENTAL STUDY

We assess the effectiveness of *VISE* for the valid link prediction (VLP) task using a synthetic Lung Cancer (SLC) KG. The VLP task aims to predict relationships, such as determining whether a patient with lung cancer is in relapse (e.g., ⟨PatientID, hasRelapse, ?⟩). The study addresses three main research questions: **RQ1)** How does KG transformation using symbolic rules and integrity constraints affect VLP accuracy? **RQ2)** How do KG size and edge variety (true/false) influence VLP performance? **RQ3)** How does VLP compare with classification tasks implemented using AI models learned from relational data?

Benchmark. We evaluate **VISE** on a synthetic Lung Cancer

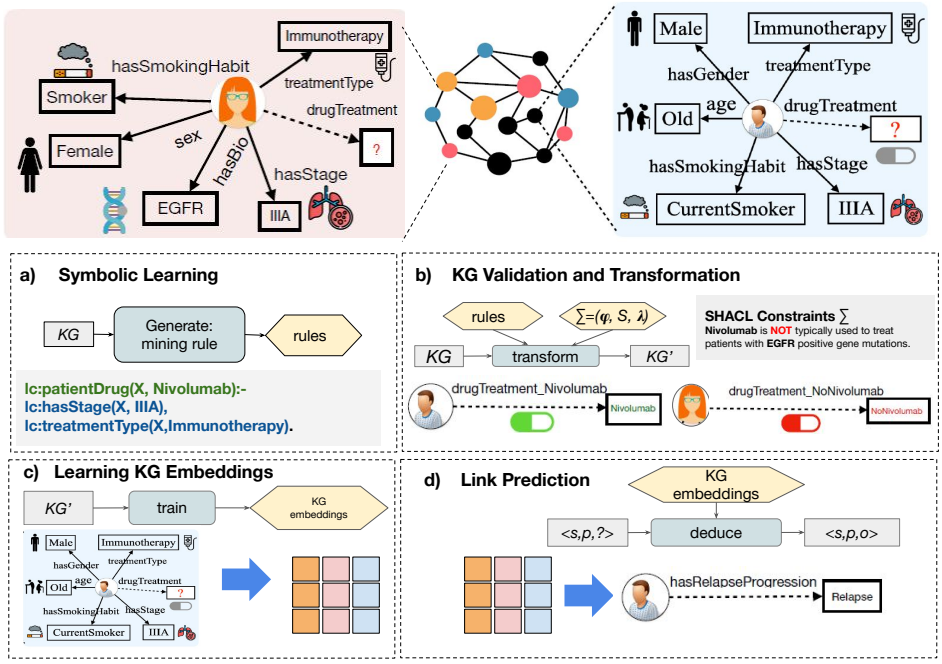


FIGURE 6: **Valid Link Prediction Example.** Figure demonstrates **VISE**, starting from symbolic learning to the link prediction step. a) Generates mining rules, shown with an example rule about prescribing Nivolumab for stage IIIA lung cancer patients receiving immunotherapy. b) Applies rules and shapes $\sum$ to transform the initial KG into KG', containing positive and negative facts. c) KG' as an input to learn embedding representation. d) Uses KGE models to predict new relationships.

KG (SLC KG) created from the KGs reported by [2]. Table 1 summarizes the statistics of the SLC KG, which contains anonymized characteristics of patients diagnosed with lung cancer. These records include medical history and characteristics such as smoking status (*e.g., non-smoker*), cancer stage (*e.g., stage IV*), demographics, cancer mutation types (*e.g., PD-L1 positive*), proposed treatments (*e.g., afatinib*), relapse status (*e.g., relapse or no relapse*), and treatment types (*e.g., chemotherapy*). The prediction problem focuses on determining the *Relapse* status of a lung cancer patient, categorized as either *Relapse* or *No Relapse*. SHACL constraints are used as medical protocols to specify when certain drugs should be prescribed based on a patient's mutations. For example, a protocol might state: *"If a patient is mutated with EGFR positive, they should not take Nivolumab"*. The SLC KG is enriched with a schema containing four SHACL constraints.

Baselines. We assess and contrast four baselines. **Baseline 1** comprises assessing the state-of-the-art KGE models for the KG completion task. **Baseline 2** shows the impact of transformed KG with KGE models. **Baseline 3** uses the hybrid approach, SPaRKLE [58], which improves the performance of KGE models by utilizing symbolic learning approaches. SPaRKLE approach is combined with patient outcomes that meet the medical protocols in **Baseline 4**. **VISE** integrates the fusion of SPaRKLE with the transformed KG including validation and violation results to enhance the performance of KGE models. The current **VISE** implementation employs various state-of-the-art KGE models from the PyKEEN [59] framework, which includes translation and rotational models such as TransE [60], TransD [61], TransH [62], and RotatE [63]. Translation-distance space models, including TransE, TransD, and TransH, translate the head entity's geometric embedding space with a given relation closer to the tail entity. A model for learning embeddings in Euclidean and hyperplane space, RotatE, has received attention for learning symmetric-and-asymmetric properties, and relationship types such as 1:1, 1:N, and M:N.

Additionally, we compare the predictive capabilities of a traditional ML i.e., Random Forests and Decision Trees [26] with KGE models in the tasks of predicting the relapse of a lung cancer patient. To assess these models, we utilize a hybrid framework, InterpretME [64], which encompasses the training of ML models and provides human-and-machine understandable interpretability of the ML models' outcomes. We consider three configurations- Without, Undersampling, and Oversampling. Without corresponds to the baseline setting where ML models are trained using the original, unmodified target class distribution, i.e., no sampling strategy is applied. In Undersampling, the strategy corresponds to the data is balanced by reducing samples from majority classes to match the number of samples in minority classes. Conversely, oversampling corresponds to achieving balance by increasing the number of samples in minority classes to match the count of majority classes. The traditional ML models are assessed in terms of Precision (P), Recall (R), and F1-score (F1).

Implementation. VISE is implemented in a virtual machine

on Google Colab with 40 GiB VRAM and 1 GPU NVIDIA A100-SMX4, with CUDA version 12.2 (Driver 535.104.05) using Python 3.10. The **VISE** implementation code, the benchmark SLC KG, and the trained KGE models are openly accessible in our GitHub repository [2]. Figure 5 demonstrates the hybrid design pattern, integrating inductive learning with symbolic learning techniques. Symbolic learning includes logical horn rules (R) and SHACL constraints (ϕ). Symbolic learning is performed over the input KG, resulting in rules, heuristic-based edges, and SHACL validation. Thus, the inferred heuristic edges with validation results are utilized as implicit knowledge to enhance inductive learning, i.e., KGE models. The predictions generated from the symbolic rules and constraints materialized in the input KG and fed as input to inductive learning. The benchmark KGs are divided into 80-20 train-test splits. The model's efficacy in the LP problem is evaluated using metrics such as *Hits@K* and *MRR* proposed by Akrami et al. [65]. Both metrics have values between 0 and 1, and higher conveys better. To avoid overfitting, the default settings for the training KGE models include a learning rate of $1e^{-1}$, and *Adam* as the regularization optimizer with a negative sampling strategy. **VISE** relies on [16] and [58] for symbolic learning methods. Furthermore, our approach is model-agnostic and compatible with other symbolic and inductive learning approaches. We employ InterpretME [33], [34], [36] to execute the traditional ML models such as Decision Trees (DT) and Random Forest (RF). InterpretME as a pipeline offers data integration, curation, and hyperparameter optimization essential for training the ML models. Moreover, in comparison to KG embeddings, the train-test split ratio is the same for traditional ML models. The predictive pipeline utilizes cross-validation (CV) [66] *k-folds* stratified shuffle split strategy, i.e., *5-folds*. The performance of the predictive models is evaluated in terms of metrics such as Precision (P), Recall(R), and F1-score (F1). Recall depicts the proportion of counts of correctly predicted patients in the RelapseProgression (RelProg) class to the total patient count with the target class Relapse in the benchmark. Precision is the ratio of accurately predicted patients in the RelapseProgression class to those projected to have class Relapse. The same evaluation parameters are used to categorize lung cancer patients as having NoRelapseProgression (No_RelProg).

D. IMPACT OF SYMBOLIC RULES AND CONSTRAINTS ON VALID LINK PREDICTION

We report the effectiveness of **VISE**, focusing on KGE models- TransE, TransD, TransH, and RotatE in the context of lung cancer relapse prediction problems. Table 2 showcases the comparison between baselines and **VISE** for link prediction. The analysis revealed a robust performance compared to baselines. KGE models are trained over the LC KG, i.e., positive edges E^+, to predict missing links. The evaluation report presented in Table 2 is obtained using the optimized hyperparameters provided by the PyKEEN pipeline. The impact of positive and negated facts is assessed with Hits@1, Hits@3, Hits@5, Hits@10, and MRR in valid link prediction. TransE, a basic translation model, emerged as performing worst in all baselines with benchmarks respectively. Nevertheless, highlighting the limitations of TransE in modeling 1:N relationships leads to poor performance, particularly in predicting the correct tail at the topmost position. TransH model results support the claim in [62], that it outperforms TransE and TransD models. In SLC KG, TransH performance contributes to promising results in capturing complex geometric relationships with Hits@1 score values ranging from 0.622 to 0.868. TransD, which uses relation-specific projections to translate the embedding space, yields slightly lower values than TransH and TransE. However, RotatE indicates the best performance in all the testbeds except in **Baseline 2**. In **Baseline 1** and **VISE**, the values of Hits@1 range from 0.696 to 0.887. We can observe that the evaluation of benchmark KG in different experimental testbeds, **VISE** outperforms compared to the other baseline approaches. The experimental evaluation comprises 20 testbeds per baseline, amounting to a total of 100 testbeds. In summary, the results of the evaluation demonstrate the effectiveness of TransH and RotatE for knowledge graph completion in the context of lung cancer relapse prediction.

However, the rationale behind the KGE models may be difficult to understand. The experimental results demonstrate the need for explanations and assistance to understand KGE model behavior. **VISE** shows improved KGE model performance and provides two types of post hoc explanation for the prediction problem. In **VISE**, KGE models showed marginally better performance compared to **Baseline 1**. The heuristic-based negative edges (hE^-) generated by symbolic learning demonstrate the effectiveness of enhancing the performance of **VISE**. The addition of hE^- edges to KG has been deemed a sufficient rationale, as evidenced by the improved performance of the KGE model in terms of Hits@K and MRR. For example, Table 5 displays examples of mined Horn rules that were chosen based on the SHACL constraints, i.e., clinical guidelines used to infer the hE^- edges. Further, the removal of these edges from KG resulted in a notable decline in performance, which can be attributed to the necessity of these facts in the prediction performance thereby answering **RQ1**).

The Horn rules are mined using AMIE [67] over SLC KG; they are selected based on biomarkers, medications, and therapies, and ranked according to the *PCA confidence* score. The effectiveness of **VISE** is evaluated in terms of the impact of validating constraints for the missing link being predicted by the symbolic learning technique. The heuristic-based negative edges (hE^-) are predicted using the Partial Completeness Assumption (PCA) heuristics from the input KG. The PCA Confidence of a Horn rule, which indicates the amount of incompleteness in a KG, is employed to infer new links and predictions. These predictions are validated by applying the SHACL constraints to determine the validity of the inferred links. The results demonstrated in Table 5 indicate the amount

[2]https://github.com/SDM-TIB/VISE

TABLE 2: **KG Evaluation.** Empirical evaluation of KGE models on SLC KG. Hits@1, Hits@3, Hits@5, Hits@10 and MRR are reported. Four baselines and **VISE** (in light green color) indicate the impact of considering semantics in prediction tasks. The values in bold convey better results and the values in underlined correspond to the best results among the baselines.

Approaches	Results for Synthetic Lung Cancer KG					
	Model	Hits@1	Hits@3	Hits@5	Hits@10	MRR
Baseline 1	**TransE**	0.000	0.560	0.795	0.943	0.324
	TransD	0.002	0.551	0.690	0.872	0.310
	TransH	0.622	0.864	0.943	0.983	0.756
	RotatE	**0.696**	**0.933**	**0.969**	**0.987**	**0.820**
Baseline 2	**TransE**	0.000	0.713	0.840	0.931	0.376
	TransD	0.008	0.694	0.824	0.935	0.379
	TransH	**0.882**	**0.969**	**0.997**	**1.000**	**0.929**
	RotatE	0.864	0.987	0.995	1.000	0.924
Baseline 3	**TransE**	0.000	0.519	0.747	0.923	0.310
	TransD	0.011	0.551	0.716	0.884	0.322
	TransH	0.596	0.876	0.925	0.977	0.740
	RotatE	**0.714**	**0.941**	**0.969**	**0.990**	**0.829**
Baseline 4	**TransE**	0.000	0.536	0.735	0.931	0.311
	TransD	0.002	0.551	0.733	0.870	0.318
	TransH	0.542	0.849	0.908	0.974	0.702
	RotatE	**0.700**	**0.945**	**0.972**	**0.992**	**0.818**
VISE	**TransE**	0.000	0.760	0.878	0.948	0.388
	TransD	0.013	0.684	0.762	0.884	0.368
	TransH	0.868	0.980	0.994	**1.000**	0.924
	RotatE	**0.887**	**0.986**	**0.996**	0.998	**0.936**

TABLE 3: **Relational Tables Evaluation.** Table displays a decision tree (DT) generated classification report for *Relapse-Progression (RelProg)* and *No RelapseProgression (NoRel-Prog)*, including Precision (P), Recall (R), and F1-score (F1).

DT	Without			Undersampling			Oversampling		
Classes	P	R	F1	P	R	F1	P	R	F1
RelProg	0.83	0.99	0.90	0.66	0.66	0.66	0.64	0.61	0.62
No_RelProg	0.57	0.04	0.09	0.64	0.65	0.65	0.64	0.66	0.65

TABLE 4: **Relational Tables Evaluation.** Table displays a random forest (RF) generated classification report for *RelapseProgression (RelProg)* and *No RelapseProgression (NoRelProg)*, including Precision (P), Recall (R), and F1-score (F1).

RF	Without			Undersampling			Oversampling		
Classes	P	R	F1	P	R	F1	P	R	F1
RelProg	0.82	1	0.90	0.68	0.71	0.70	0.67	0.71	0.69
No_RelProg	0.00	0.00	0.00	0.70	0.67	0.68	0.70	0.65	0.67

of valid and invalid predictions produced by the symbolic learning techniques. Table 5 shows examples of the symbolic rules, for example, `stage(?a, IV), treatment(?a, Immunotherapy)` $\Rightarrow$ `drug(?a, Nivolumab)` stating that if a stage IV lung cancer patient received *Immunotherapy* treatment then it is more likely that the patient receives Nivolumab is with the *PCA Confidence* score of 0.83. As mentioned before, the heuristics-based negative edges (hE^-) or predictions are validated using SHACL constraints, and Table 5 shows the number of valid ($\#v$) and invalid ($\#in$) links for each of the SLC KG used as a benchmark in **VISE**. Consequently, the impact of symbolic rules and constraints utilized to explain the KGE models is demonstrated, thereby enabling an answer to be provided to the **RQ2**.

E. TRADITIONAL ML VERSUS KGE MODELS

Experimental evaluation of traditional ML models and KG embeddings reveals explicit complex patterns of performance and predictive capabilities. We aim to evaluate the research question (**RQ3**) in this study. Tables 3 and 4 show the classification outcomes for Decision Trees and Random Forests in three configurations - Without, Undersampling, and Oversampling. In traditional ML, Table 3 shows Decision Trees (DT) performed best with a precision of 0.83 and a promising recall of 0.99 (F1 score: 0.90) for relapse progression prediction under standard settings, while Table 4 shows Random Forests (RF) showed comparable precision (0.82) and perfect recall (1.00) for similar cases. RF also struggled with no-relapse progression predictions, suggesting limitations in handling class imbalance distributions.

In KG, embedding models evaluated through **VISE** and multiple baselines (Table 2) showed different performance patterns, with RotatE emerging as a consistently strong performer, achieving impressive metrics (Hits@1: 0.887, Hits@3: 0.986, Hits@5: 0.996, Hits@10: 0.998, MRR: 0.936 under **VISE**). The overlap analysis of correctly predicted links in Figure 7 reveals an intriguing complementarity between different inductive learning approaches, with RotatE showing the strongest alignment with decision trees (83.6% overlap, 219 shared predictions), suggesting its ability to capture both traditional feature-based patterns and complex relationships. TransE and RotatE showed substantial prediction alignment (72.5%, 190 shared predictions), indicating RotatE's ability to preserve translational relationships while adding rotational modeling power. The lower overlap between TransE and TransD (54.2%, 142 predictions) indicates that these models capture disparate relationships in SLC KG.

VISE enhances performance across most KGE models, particularly benefiting TransH and RotatE, with TransH achieving perfect Hits@10 scores and strong MRR (0.924). These

TABLE 5: **Exemplary Mined Horn Rules.**

Exemplary Mined Horn Rules	Natural Language Statements	ORKG ASK	PCA Conf.	hE #v	hE #in
drug(?a, Nivolumab)⇐ stage(?a, IV) treatment(?a, Immunotherapy)	IF a Lung Cancer patient is in cancer stage IV and receives treatment type as immunotherapy THEN a patient could receive Nivolumab drug	Ask 1	0.83	50	40
biomarker(?a, EGFR Negative)⇐ relapseProgression(?a, Progression) drug(?a, Pembrolizumab)	IF a Lung Cancer patient is in progression and receives Pembrolizumab drug THEN a patient has EGFR Negative mutation	Ask 2	0.97	0	10
drug(?a, Nivolumab)⇐ treatment(?a, Immunotherapy), treatment(?a, Intravenous Chemotherapy)	IF a Lung Cancer patient receives treatment type such as immunotherapy and Intravenous Chemotherapy THEN a patient could receive Nivolumab drug	Ask 3	0.72	130	100
biomarker(?a, EGFR Negative)⇐ biomarker(?a, ALK Negative), treatment(?a, Radiotherapy to Bone)	IF a Lung Cancer patient has ALK Negative and receives treatment type as Radiotherapy to Bone THEN a patient has EGFR Negative mutation	Ask 4	0.92	20	20

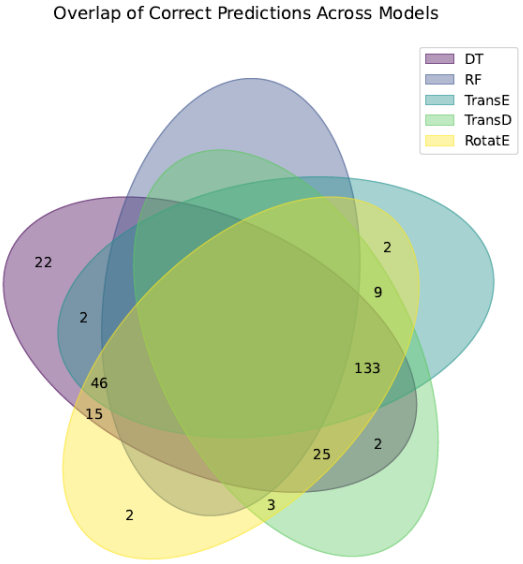


FIGURE 7: **Overlap Analysis.** Different models for several predictions; the analysis includes: Decision Trees (DT), Random Forest (RF), TransE, TransD, TransH.

patterns suggest that while individual models excel in specific aspects, they offer complementary strengths: traditional ML models provide interpretable predictions for straightforward cases, while KGE models, especially RotatE, excel at capturing complex relationships in the KG structure. The varying overlaps despite strong individual performances indicate that different models capture distinct aspects of the underlying relationships, supporting the potential value of ensemble approaches. TransH and RotatE's superior performance under **VISE**, combined with RotatE's strong overlaps with other models, suggests these approaches might serve as robust foundations for predictive models that require both relationship modeling capability and traditional feature-based prediction power.

F. DISCUSSION AND LESSONS LEARNED

The observed results indicate that existing KG embedding (KGE) models are significantly influenced by how factual statements are represented in KGs, with performance improvements observed when valid and invalid links are explicitly defined according to domain-specific integrity constraints. This evidence supports that the hybrid approach implemented in **VISE** effectively leverages the semantics of integrity constraints and mined rules to improve the predictive accuracy of KGE models, providing a more precise solution to the valid link prediction problem. Additionally, as described in subsection III-A, TrustKG facilitates human-machine communication by translating mined rules into natural language and retrieving relevant publications from ORKG ASK. Table 5 presents examples of mined rules, natural language explanations, and related scientific papers. These results also highlight the role of the hybrid methods in **VISE** in improving the interpretability of its results.

V. COUNTERFACTUAL PREDICTION

This section defines the problem of counterfactual and presents **HealthCareAI**. The boxology of design patterns proposed by Van Bekkum et al. [57] is also used to specify the main components of **HealthCareAI**. We start presenting basic concepts like *causal analysis*, *causal relationships*, and *causal models*.

Causal Analysis. Causal analysis in KGs examines the effects of interventions or treatments on outcomes. An entity in the KG, known as a *unit*, undergoes an intervention represented by $do(T = t)$, where T is the treatment variable and t is a specific treatment value. It distinguishes between observed (factual) outcomes and hypothetical (counterfactual) outcomes under alternative treatment scenarios, enabling reasoning about cause-and-effect relationships in KG.

Causal Relationships. Causal relationships in KGs are represented using several models: i) *Causal Graphs (CGs):* Directed acyclic graphs (DAGs) that depict direct causal relationships between variables. In a causal graph $G = (X, E^c)$, a directed edge $(X_i, X_j) \in E^c$) indicates that X_i has a causal influence on X_j. ii) *Causal Bayesian Networks (CBNs):* Extending causal graphs, CBNs encode probabilistic relationships and intervention effects among variables. The joint probability distribution $P(X)$ over X can be factored based on the graph structure. When an intervention $do(T = t)$ occurs, the interventional distribution $P_t(X)$ computed adjusting the influence of other variables on T.

Causal Models. Causal models in KGs map treatments and contexts to outcomes, predicting how interventions affect specific entities. For a KG $KG = (V, E, L)$, key concepts include: i) A *target class* C, representing the type of entities (e.g., patients) being analyzed. ii) A *treatment property* p_T and an *outcome property* p_Y associated with class C. iii) *Contextual properties* P_C^{ctx}, representing additional factors related to C that may influence the outcome.

For an entity e of class C, the treatment is defined by (e, p_T, t) and the outcome by (e, p_Y, y). The context consists of all the triples involving properties in P_C^{ctx}. A causal model ϑ maps a treatment and its context to an outcome as $\vartheta((s, p_T, t), E_{ctx}) = (s, p_Y, y)$, where E_{ctx} contains the contextual information for s. The unit dataset $D_{KG}(C)$ aggregates entities, treatments, outcomes, and contexts of type C, enabling causal effect estimation across similar units.

A. PROBLEM STATEMENT

Given a Knowledge Graph *KG = (V, L, E)*, a target class C with its properties P_C, including treatment p_T, outcome p_Y, and contextual properties P_C^{ctx}, the goal of counterfactual prediction is to determine the outcome for hypothetical treatments. Let $V_t = \{o \mid (s, p_T, o) \in E\}$ represent possible treatment values. For a unit e of type C, we consider a hypothetical treatment (e, p_T, t') where $t' \in V_t$ but $(e, p_T, t') \notin E$. The objective is to find an optimal *causal model* ϑ^* for predicting the outcome p_Y under this hypothetical treatment. This optimal model ϑ^* is selected from the space of all possible causal models Θ to maximize a utility function $f(.,.)$ that measures the accuracy of the counterfactual prediction:

$$\boxed{\vartheta^* = \underset{\vartheta \in \Theta}{\operatorname{argmax}} f(\vartheta((e, p_T, t'), \phi(e, P_C^{ctx})))}$$

$f(.,.)$ evaluates how accurately a causal model ϑ predicts the counterfactual outcome based on the treatment and context.

B. PROPOSED SOLUTION

HealthCareAI solves the problem of counterfactual prediction by performing three main tasks: a) Symbolic reasoning; b) Causal model learning; c) Counterfactual prediction (shown in Figure 8). *Symbolic Reasoning* The input (Fig. 8-**a**) includes a knowledge graph $KG = (V, E, L)$, a set of Horn rules R over the properties in L, a target class $C \in V$, and treatment and outcome properties p_T and p_Y in L. The set R consists of logical rules, such as those defined in `RDFS` [68] or `OWL` [69]. The transformation of KG is guided by a semantic model and an axiomatic system that states the meaning of symbols in KG. Query processing and symbolic learning are performed over KG, applying the logical rules in R to generate an enriched KG' that includes deduced facts and domain knowledge [45]. Domain knowledge includes: i) Metadata for the target class C and its properties $P_C \subseteq L$ (e.g., `rdfs:label`, `rdfs:domain`) [70]. ii) A set of additional Horn rules R', mined through symbolic learning, which meet certain criteria, such as a minimum confidence and PCA confidence score [58]. These mined Horn rules, which may imply causal relationships, form part of the domain knowledge derived from KG'. This knowledge is used to construct a prompt for a large language model (LLM) [71] to query causal relationships between the properties in P_C. Additionally, KG' is used to create a unit dataset $D_{KG'}(C)$, which supports further causal model learning.

Causal Model Learning (Fig. 8-**b**) component infers a CG (for the causal model ϑ) over the properties P_C using a hybrid approach. It takes as input the unit dataset $D_{KG'}(C)$ and the LLM prompt from the previous component, and generates a casual model ϑ for counterfactual prediction. To do this, it combines a data-driven statistical model and a knowledge-driven model LLM to infer CG. The statistical model learns a CG $G_1 = (P_C, E_1)$ directly from the dataset $D_{KG'}(C)$. The statistical model can be any traditional causal discovery method, such as **PC** [72], **FCI** [72], or **GES** [46]. The metadata-driven model (any LLM) takes as input an LLM prompt and returns a CG $G_2 = (P_C, E_2)$. The LLM prompt is designed into four sections, where the role section uses the metadata to provide the domain information of KG and specify the functions of the LLM. The context section uses the metadata of the target class C and properties P_C, including the domain (`rdfs:domain`) and range (`rdfs:range`) of the properties, the human-readable label (`rdfs:label`) and annotation (`rdfs:comment`). In addition, each Horn rule in R' is translated into a set of association pairs (p_1,p_2), where p_1 and p_2 are the predicates in the head and body of a Horn rule, respectively, as potential causal relationship candidates; the objective section specifies the task of identifying causal relationships between properties in P_C; the instruction section formats the output of causal relations. The final output of this component is a CG $G = (P_C, E^c)$ s.t. $E^c = E_2 \cup \{(c, e) \in E_1 | (P_C, E_2 \cup \{(c, e)\})$ is a DAG$\}$. In other words, it includes all causal relations in E_2 (by LLM) and those in E_1 that do not introduce any directed circle in G. The causal model ϑ is trained based on G and $D_{KG'}(C)$.

Counterfactual Prediction (see Fig. 8-**c**) predicts counterfactuals from units. Given a causal model ϑ, a hypothetical treatment (e, p_T, t') and context $\phi(e, P_C^{ctx})$ of a unit e, it predicts the counterfactual on property p_Y of e. As a proof of concept, we use CBN as the *causal model* ϑ. Let $\vartheta : (G, P_t)$ be the CBN based on the CG G trained on the dataset $D_{KG'}(C)$. The counterfactual p_Y of e under a hypothetical treatment $(e, p_T, t') \notin E$ is predicted as (e, p_Y, y') s.t. $P_{t'}(y'|x) \geq max(\{P_{t'}(y|x)\}_{y \in V_Y})$, where $V_Y = \{y | (\exists e' \in V)[(e', p_Y, y) \in E]\}$ contains all unique values of p_Y, and $P_t(y|c) = \frac{\sum_s P_t(y,c,s)}{\sum_{y,s} P_t(y,s)}$ is an interventional distribution (see the definition of CBNs) derived from the CBN model ϑ.

To illustrate, consider a lung cancer patient, *Eva* (depicted on the right side of Fig. 8-**a**), a non-smoking female with a positive ALK biomarker result. The objective is to predict her biomarker result under the counterfactual scenario where she is a smoker (shown in Fig. 8-**c**). As a result of symbolic reasoning, **HealthCareAI** extracts domain knowledge from KG, including two parts: (1) metadata stored within KG, and

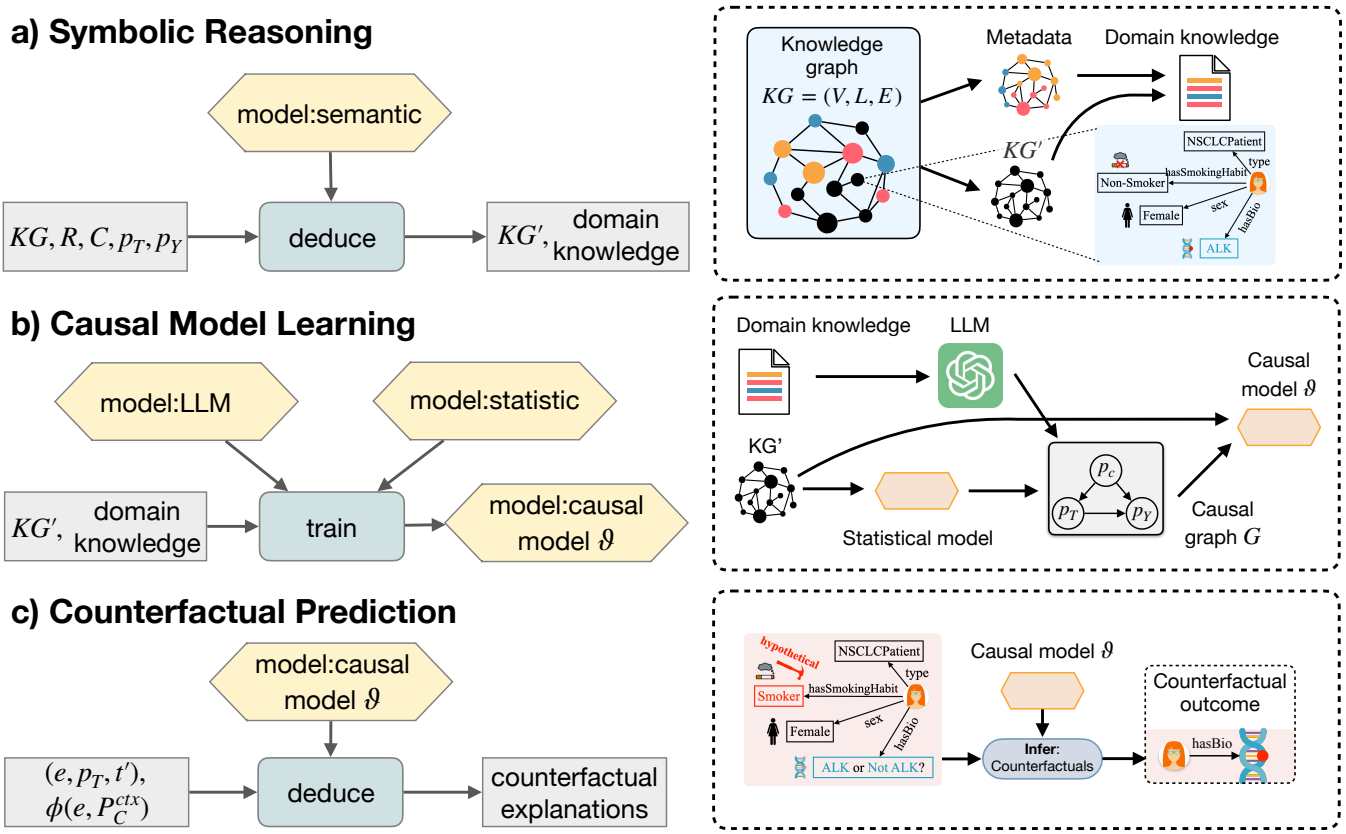


FIGURE 8: **HealthCareAI framework (a) Symbolic Reasoning:** A semantic model deduces an enriched knowledge graph KG' and domain knowledge (metadata and Horn rules) from the input KG, Horn rules R, target class C, and properties p_T and p_Y. **(b) Causal Model Learning:** A causal model ϑ is trained by integrating a statistical model (inferring causal graph G_1 from $D_{KG'}(C)$) and an LLM (producing causal graph G_2 from prompts). The combined graph G ensures acyclicity and supports causal inference. **(c) Counterfactual Prediction:** Using ϑ, hypothetical treatments (e, p_T, t'), and contextual information $\phi(e, P_C^{ctx})$, counterfactual outcomes p_Y are predicted, providing interpretable explanations. The framework integrates symbolic and statistical learning for accurate and interpretable counterfactual reasoning.

(2) Horn rules mined from an enriched KG' that incorporates implicit facts inferred from KG using a formal system. During the causal model learning phase (see Fig. 8-**b**), an LLM utilizes the domain knowledge to deduce causal relationships between the properties of the target class—in this case, the `NSCLCPatient`. In addition, a statistical model directly learns causal relationships from KG'. The inferred relationships are integrated into a single causal graph G using the method specified in the causal model learning component. Subsequently, the causal model ϑ is learned using the data from KG' and G. Finally, the causal model ϑ receives the contextual properties of *Eva* and a hypothetical treatment (i.e., (`Eva`, `hasSmokingHabit`, `Smoker`)) and infers the counterfactual biomarker result of *Eva* if she were a smoker.

C. EXPERIMENTAL STUDY

We evaluate the effectiveness of **HealthCareAI** in causal discovery and counterfactual prediction tasks over a synthetic LC KGs with different sizes; these benchmarks are generated from the LC KG reported by Calvo et al. [2]. The study is

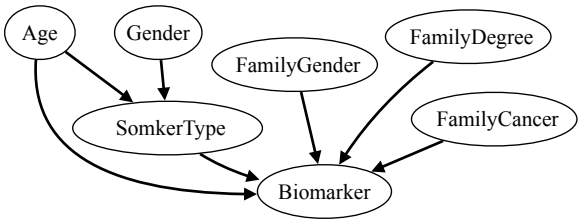


FIGURE 9: **Expert** Designed Causal Graph G^* built over the properties in synthetic Lung Cancer KGs

guided by the following research questions:

RQ4 How does the neuro-symbolic system improve the discovery of causal relations as defined by the expert causal graph?

RQ5 How does the neuro-symbolic system influence the accuracy and effectiveness of counterfactual predictions?

Benchmarks. The synthetic KGs include the following properties: `Biomarker` that is either *ALK or EGFR* and *other biomarkers*, `Age` that is categorized as *Young* ($\leq$ *50 years)* or *Old* ($>$ *50 years)*, `Gender` that is *Male*

TABLE 6: Benchmarks of synthetic KGs for counterfactual prediction. **#CF** represents the number of ground truth counterfactuals (triples).

KG	KG_{2k}	KG_{5k}	KG_{10k}
#triples	16180	40180	80180
#entities	2055	5055	10055
#relations	17	17	17
#CF	2000	5000	10000

or *Female*, `SmokerType` that is *Non-Smoker* or *Smoker*, `FamilyCancer` which is *OnlyMajor* if a patient's family antecedents have only these cancers: *Breast*, *Lung*, *Colorectal*, *Head and neck*, *Uterus/cervical*, *Esophagogastric*, *Prostate*, otherwise *hasMinor*, `FamilyGender` that is *Women* if all family antecedents are women, *Men* if all of them are men, otherwise *WomenorMen*, and `FamilyDegree` that indicates the degree of relationship of family antecedents to the patient (i.e., first, second, or third degree). We train an Additive Noise Model (ANM) [73] using the expert-designed CG G^* (see Figure 9), using the LC KG [2]. Using this trained ANM model (assumed to be able to capture the causal mechanism implied in the dataset), we generate synthetic KGs with various number of patients $N \in \{2k, 5k, 10k\}$. Specifically, `Age`, `Gender`, `FamilyGender`, `FamilyCancer` and `FamilyDegree` are simulated from uniform distributions; while `SmokerType`, and `Biomarker` are simulated using logistic functions:

$$Y = 1/(1 + exp(-(\alpha + \beta \cdot X' + N(0, \sigma^2)))) \quad (1)$$

where Y is the synthetic variable; $X' = Pa_{G^*}(Y)$ include the parent nodes of Y in the CG G^* (see Figure 9). The α and β are learned from the LC KG using a logistic regression. A noise term $\mathcal{N}(0, 0.1^2)$ is applied to simulate other potential unobserved factors. Additionally, each synthetic KG includes the metadata of the LC KG. To evaluate the performance of counterfactual prediction, we generate the ground truth counterfactuals on `Biomarker` under intervention on `SmokerType` for each patient using the trained ANM. Given a synthetic $KG = (V, E, L)$, a patient $e \in V$ whose treatment is $(e, p_T, t) \in E$. Let Y=1 denote that `Biomarker` (p_Y) is *ALK or EGFR*, otherwise (Y=0) is *other biomarkers*, and T=t denotes the treatments on `SmokeType` (p_T) which are *Non-Smoker* (t=1) and *Smoker* (t=0). The counterfactual `Biomarker` of a patient (unit) e is generated using the Function 1 with input of $X' = x'$ following the assignment $X'_1, \ldots, X'_k := x'_1, \ldots, x'_k$ s.t. $(\forall i \in [1, k])[(e, X'_i, x'_i) \in S]$ where $k = |X'|$ is the variable number of X' and S is a set of triples $\phi(e, P_C^{ctx}) \cup \{(e, p_T, t')\}$ s.t. $(e, p_T, t') \notin E$. Table 6 presents the statistics of the synthetic LC KGs, denoted as KG_{2k}, KG_{5k}, and KG_{10k}.

Baselines. We compare **HealthCareAI** with other baselines, including **Baseline1**: methods using only statistical models (such as **PC**, **FIC**, and **GES**) for causal discovery, **Baseline2**: Methods that use only LLM to query the causal relationships without (**wo**) consider domain knowledge (**DK**), **Baseline3**: Methods that uses LLM to query the causal relationships with (**w**) consider the domain knowledge, **Baseline4**: the method using the expert-designed CG G^* (in Fig. 9) to learn the causal model ϑ.

Metrics. For the causal discovery task, we evaluate an inferred CG G against the expert-designed CG G^* (in Fig. 9), using the Jaccard Index [74], Precision, Recall, and F1-score [75] to measure the proportion of shared causal relationships between G and G^*. For the counterfactual prediction task, we evaluate the predicted counterfactuals against the ground truth. Given a list of predicted counterfactuals $\hat{Y}^{CF}$ by a CBN and the ground truth counterfactuals Y^{CF}, we calculate the *Pearson correlation coefficient* [76] (PCC): $PCC(\hat{Y}^{CF}, Y^{CF}) = \frac{cov(\hat{Y}^{CF}, Y^{CF})}{\sigma_{\hat{Y}^{CF}} \sigma_{Y^{CF}}}$, where $cov(\cdot)$ and σ represent covariance and standard deviation. Higher values indicate better performance.

Implementation. We implement the set of logical rules R (see Fig. 8-**a**) as an empty set, since each patient in a synthetic KG has completed properties. The Horn rules R', as a part of the domain knowledge, are mined from the synthetic KGs using the AMIE [67]. We implement the LLM by the GPT-4 [77]. As proof of concept, we use CBNs as the causal model ϑ (see Fig. 8-**b**) for all methods. CBNs of all methods are implemented using `pgmpy` Python package [78]. The reproduction code and synthetic datasets are available here [3], where the LLM prompts used by **Baseline2**, **Baseline3**, and **HealthCareAI** are presented.

D. IMPACT ON CAUSAL DISCOVERY

We report the results on causal discovery task in Table 7 (for synthetic Lung Cancer KG with $N = 2k$, denoted as KG_{2k}), Table 8 (for $N = 5k$, denoted as KG_{5k}), and Table 9 (for $N = 10k$, KG_{5k}). In general, all methods have better precision performance but relatively weaker recall. This means that they are able to correctly uncover some causal relationships in CG designed by experts G^* (see Figure 9), but the inferred causal edges are not complete. Among, all methods, the **HealthCareAI**, which combines **Baseline3** and **PC** method, outperforms other methods in all synthetic KGs with different settings of N. The data-driven methods, i.e., methods of **Baseline1**, have similar performance compared with the metadata-driven methods (**Baseline2** and **Baseline3**) in small KG (with $N = 2k$); their performance is improved in larger KGs with $N \in \{5k, 10k\}$. The **PC** and **FCI** have exact the same performance across all KGs. This may be explained that both are based on the conditional independence test and there are no hidden confounders in our simulation where **FCI** is more suitable. Even the metrics of **PC** (or **FCI**) are the same in KG_{5k} and KG_{10k}, while the inferred CGs in both situations are not exactly the same. Among all methods in **Baseline1**, the **GES** performs the worst. For metadata-driven methods, i.e., **Baseline2** and **Baseline3**, the performance is stable in all settings of N, because these methods do not depend on data. The results of **Baseline3** is constantly better

[3] https://github.com/SDM-TIB/HealthCareAI

TABLE 7: Comparison of Causal Discovery Performance of Different Approaches Against the Expert Causal Graph. **w** and **wo** denote respectively "with" and "without", and **DK** denotes "domain knowledge".

Approaches	Results (%) for KG_{2k}				
	Model	Jaccard Index	Precision	Recall	F1-Score
Baseline 1	**PC**	62.5	**100.0**	62.5	76.9
	FCI	62.5	**100.0**	62.5	76.9
	GES	37.5	**100.0**	37.5	54.5
Baseline 2	**GPT4 wo DK**	40.0	66.7	50.0	57.1
Baseline 3	**GPT4 w DK**	62.5	**100.0**	62.5	76.9
HealthCareAI	**Baseline 2 + PC**	60.0	75.0	75.0	75.0
	Baseline 3 + PC	**87.5**	**100.0**	**87.5**	**93.3**

TABLE 8: Comparison of Causal Discovery Performance of Different Approaches Against the Expert Causal Graph. **w** and **wo** denote respectively "with" and "without", and **DK** denotes "domain knowledge".

Approaches	Results (%) for KG_{5k}				
	Model	Jaccard Index	Precision	Recall	F1-Score
Baseline 1	**PC**	75.0	**100.0**	75.0	85.7
	FCI	75.0	**100.0**	75.0	85.7
	GES	62.5	**100.0**	62.5	76.9
Baseline 2	**GPT4 wo DK**	40.0	66.7	50.0	57.1
Baseline 3	**GPT4 w DK**	62.5	**100.0**	62.5	76.9
HealthCareAI	**Baseline 2 + PC**	70.0	77.8	**87.5**	82.4
	Baseline 3 + PC	**87.5**	**100.0**	**87.5**	**93.3**

TABLE 9: Comparison of Causal Discovery Performance of Different Approaches Against the Expert Causal Graph. **w** and **wo** denote respectively "with" and "without", and **DK** denotes "domain knowledge".

Approaches	Results (%) for KG_{10k}				
	Model	Jaccard Index	Precision	Recall	F1-Score
Baseline 1	**PC**	75.0	**100.0**	75.0	85.7
	FCI	75.0	**100.0**	75.0	85.7
	GES	62.5	**100.0**	62.5	76.9
Baseline 2	**GPT4 wo DK**	40.0	66.7	50.0	57.1
Baseline 3	**GPT4 w DK**	62.5	**100.0**	62.5	76.9
HealthCareAI	**Baseline 2 + PC**	70.0	77.8	87.5	82.4
	Baseline 3 + PC	**100.0**	**100.0**	**100.0**	**100.0**

than **Baseline2** which demonstrates the usefulness of domain knowledge (**DK**) in causal discovery. The performance of data-driven methods (**Baseline1**) improves as the data size N increases, but peaks at $5k$ and $10k$. In addition, learning the CG from (observational) data alone may be theoretically impossible [40]. As a complement, the metadata- or knowledge-driven methods (such as **Baseline3**) can infer some causal relationships that cannot revealed by **Baseline1** methods, this explains why **HealthCareAI** can achieve the best performance in all synthetic KGs. These results answer the research question **RQ4** that the neurosymbolic systems enriched with semantics can improve the performance of traditional data-driven methods by using knowledge deduced from KGs.

E. IMPACT ON COUNTERFACTUAL PREDICTION

Based on the CGs estimated from the previous step (subsection V), we learn CBNs (causal model ϑ) from the dataset $D_{KG'}(C)$ over the synthetic LC KGs using the *Maximum Likelihood Estimation* Method [79]. The evaluation results of different methods on counterfactual prediction in various synthetic LC KGs with $N \in \{2k, 5k, 10k\}$ are reported, respectively, in Table 10, Table 11, and Table 12. The PCC metrics for the CBN by each method are presented as the mean ($\pm$ standard deviation), using 5-fold cross-validation on all patients and their counterfactuals. The results indicate that **PC** and **FCI** of **Baseline1** outperform others in scenarios with smaller datasets (i.e., KG_{2k}). In contrast, the CBN trained based on the **Expert** CG G^* exhibits the lowest performance, which may be explained by the overfitting issue [80]. This is likely due to the complex structure of the **Expert** CG, which requires a large dataset to learn the conditional probability tables of the CBN. Conversely, the simpler structures of the CGs deduced by the **PC** and **FCI** methods allow for effective learning with small datasets. In the scenario of large datasets (i.e., $N \geq 5k$), the Expert CBN outperforms all CBNs trained based on CGs by other methods. Although enlarging the dataset generally enhances the generalization capability of CBNs, it is notable that the performance of all CBNs slightly declines as N increases from $5k$ to $10k$. The CBNs by the metadata-driven methods, i.e., the **Baseline2** and **Baseline3** perform the worst across all synthetic KGs. This result underscores the limitations of CGs estimated without considering the underlying data, emphasizing the crucial role of data-driven causal graph estimation for robust counterfactual reasoning. By incorporating the causal relationships estimated by the data-driven methods, e.g., **PC**, the CBN produced by **HealthCareAI** achieves competitive performance across all settings of N, compared to the CBN based on the **Expert** CG. The results address the research question **RQ5**, confirming that **HealthCareAI** outperforms the other baselines and exhibits a good counterfactual predicting performance with respect to **Baseline4**, i.e., the one based on the Expert CBN.

TABLE 10: Comparison of Counterfactual Prediction using Causal Models based on Different Causal Graphs produced by different Approaches at $N = 2k$. **Baseline4** use the expert-designed CG G^*. **w** and **wo** denote respectively "with" and "without", and **DK** denotes "domain knowledge".

Approaches	Results (%) for KG_{2k}	
	Model	Pearson Correlation
Baseline 1	**PC**	**95.0 (±2.5)**
	FCI	**95.0 (±2.5)**
	GES	89.9 (±1.6)
Baseline 2	**GPT4 wo DK**	89.9 (±1.6)
Baseline 3	**GPT4 w DK**	87.0 (±1.8)
Baseline 4	**Expert causal graph**	81.4 (±2.9)
HealthCareAI	**Baseline 2 + PC**	93.9 (±2.7)
	Baseline 3 + PC	81.4 (±2.9)

TABLE 11: Comparison of Counterfactual Prediction using Causal Models based on Different Causal Graphs produced by different Approaches at $N = 5k$. **Baseline4** use the expert-designed CG G^*. **w** and **wo** denote respectively "with" and "without", and **DK** denotes "domain knowledge".

Approaches	Results (%) for KG_{5k}	
	Model	Pearson Correlation
Baseline 1	**PC**	95.9 (±0.8)
	FCI	95.9 (±0.8)
	GES	95.9 (±0.8)
Baseline 2	**GPT4 wo DK**	90.7 (±1.6)
Baseline 3	**GPT4 w DK**	90.1 (±1.3)
Baseline 4	**Expert causal graph**	**96.4 (±0.7)**
HealthCareAI	**Baseline 2 + PC**	**96.4 (±0.7)**
	Baseline 3 + PC	**96.4 (±0.7)**

TABLE 12: Comparison of Counterfactual Prediction using Causal Models based on Different Causal Graphs produced by different Approaches at $N = 10k$. **Baseline4** use the expert-designed CG G^*. **w** and **wo** denote respectively "with" and "without", and **DK** denotes "domain knowledge".

Approaches	Results (%) for KG_{10k}	
	Model	Pearson Correlation
Baseline 1	**PC**	95.1 (±0.8)
	FCI	95.1 (±0.8)
	GES	95.1 (±0.8)
Baseline 2	**GPT4 wo DK**	90.0 (±0.9)
Baseline 3	**GPT4 w DK**	89.5 (±0.6)
Baseline 4	**Expert causal graph**	**95.9 (±0.6)**
HealthCareAI	**Baseline 2 + PC**	95.1 (±0.8)
	Baseline 3 + PC	**95.9 (±0.6)**

F. DISCUSSION AND LESSONS LEARNED

These results show the benefits of integrating semantics and domain-specific knowledge into the causal learning pipelines. By leveraging metadata and mined rules from KGs, **HealthCareAI** improves both the accuracy and interpretability of causal models. Semantic information enables a more structured representation of causal relationships, which not only enhances predictive capabilities but also aligns more closely with human reasoning processes. This hybrid approach also underscores the importance of using enriched domain knowledge to learn accurate causal graphs, supporting detailed counterfactual predictions. Finally, these findings suggest that the explicit representation of semantics within causal models contributes to better decision-making and will facilitate human-machine communication by making results more interpretable and contextually grounded.

VI. CONCLUSIONS AND FUTURE WORK

This work has demonstrated the potential of integrating semantics within hybrid AI systems to enhance predictive capabilities, contextual understanding, and interpretability in medical applications, particularly for lung cancer. Our proposed framework, TrustKG, along with the hybrid AI systems **VISE** and **HealthCareAI**, leverages KGs and symbolic reasoning to address critical tasks such as valid link prediction and counterfactual prediction. The results show that semantic integration improves predictive accuracy, as evidenced by the enhanced Hits@1 and MRR scores achieved by **VISE**, and supports more accurate causal discovery, with **HealthCareAI** outperforming baselines on metrics like the Jaccard Index and F1-Score. Additionally, the use of counterfactual reasoning in **HealthCareAI** achieves Pearson Correlation scores close to expert-derived benchmarks, illustrating the effectiveness of integrating domain-specific knowledge. These findings highlight the role of semantic knowledge in advancing AI systems for healthcare by providing more transparent solutions that align with clinical needs.

Our future directions include:

Evolving Data Management: One key area is enhancing semantic data management to model evolving and non-monotonic knowledge. This includes handling changes in medical guidelines, emerging treatments, and patient-specific

factors over time. Improved models for representing these dynamic knowledge can support AI systems that remain adaptable in clinical practice.

Reducing Computational Costs: While effective, our approach increases execution costs. Future work should focus on optimizing the computational efficiency of these models, potentially through more efficient embedding methods or streamlined symbolic reasoning processes.

Principled vs. Integrated Neuro-Symbolic Systems: There is a need to explore both principled and integrated approaches to neuro-symbolic AI. Principled systems maintain a clear separation between symbolic and neural components, while integrated systems combine these elements more fluidly. Research into the advantages and limitations of each approach can inform the development of hybrid AI systems that best meet clinical needs.

Enhancing Usability through Better Visualization and User Interfaces: Usability remains a critical challenge in deploying KG-based AI systems in clinical settings. Enhanced visualization techniques, as well as intuitive interfaces, can improve user experience and help clinicians interpret AI-generated recommendations more effectively. Additionally, understanding user needs, as highlighted by recent research, will guide the development of tools that align with clinical workflows and support decision-making processes. These future directions aim to refine the interpretability, usability, and efficiency of hybrid AI systems in healthcare, making them more practical in real-world applications. The integration of semantics within AI holds substantial promise, not only for advancing predictive accuracy but also for fostering trustworthy and human-centered AI solutions in medicine.

ACKNOWLEDGEMENTS

This work has been partially supported by the Leibniz Association in the program "Leibniz Best Minds: Programme for Women Professors", project TrustKG-Transforming Data in Trustable Insights with grant P99/2020.

REFERENCES

[1] Claudio Gutiérrez and Juan F Sequeda. Knowledge graphs. *Communications of the ACM*, 64(3):96–104, 2021.

[2] Virginia Calvo, Emetis Niazmand, Enric Carcereny, Delvys Rodriguez-Abreu, Manuel Cobo, Rafael López-Castro, María Guirado, Carlos Camps, Ana Laura Ortega, Reyes Bernabé, Bartomeu Massutí, Rosario Garcia-Campelo, Edel del Barco, José Luis González-Larriba, Joaquim Bosch-Barrera, Marta Martínez, María Torrente, María-Esther Vidal, and Mariano Provencio. Family history of cancer and lung cancer: Utility of big data and artificial intelligence for exploring the role of genetic risk. *Lung Cancer*, 195:107920, 2024.

[3] F. Aisopos et al. Knowledge graphs for enhancing transparency in health data ecosystems. *Semantic Web*, 2023.

[4] Maria-Esther Vidal, Kemele M. Endris, Samaneh Jazashoori, Ahmad Sakor, and Ariam Rivas. Transforming heterogeneous data into knowledge for personalized treatments - A use case. *Datenbank-Spektrum*, 19(2):95–106, 2019.

[5] Alex Molassiotis, Patsy Yates, and Janelle Yorke. Editorial: Quality of life and side effects management in lung cancer treatment. *Frontiers in Oncology*, 11, March 2021.

[6] Satya Prakash Maurya, Pushpendra Singh Sisodia, Rahul Mishra, and Devesh Pratap singh. Performance of machine learning algorithms for lung cancer prediction: a comparative approach. *Scientific Reports*, 14(1), August 2024.

[7] Zahra Sadeghi, Roohallah Alizadehsani, Mehmet Akif CIFCI, Samina Kausar, Rizwan Rehman, Priyakshi Mahanta, Pranjal Kumar Bora, Ammar Almasri, Rami S. Alkhawaldeh, Sadiq Hussain, Bilal Alatas, Afshin Shoeibi, Hossein Moosaei, Milan Hladík, Saeid Nahavandi, and Panos M. Pardalos. A review of explainable artificial intelligence in healthcare. *Computers and Electrical Engineering*, 118:109370, 2024.

[8] Samuele Buosi, Mohan Timilsina, Adrianna Janik, Luca Costabello, Maria Torrente, Mariano Provencio, Dirk Fey, and Vít Nováček. Machine learning estimated probability of relapse in early-stage non-small-cell lung cancer patients with aneuploidy imputation scores and knowledge graph embeddings. *Expert Syst. Appl.*, 235:121127, 2024.

[9] Disha Purohit, Yashrajsinh Chudasama, Maria Torrente, and Maria-Esther Vidal. Vise: Validated and invalidated symbolic explanations for knowledge graph integrity. In *CEUR Proceedings of the First Workshop on Explainable Artificial Intelligence for the Medical Domain (EXPLIMED 2024), co-located with the 27th European Conference on Artificial Intelligence (ECAI 2024)*. CEUR-WS.org, 2024.

[10] Hao Huang, Emetis Niazmand, and Maria-Esther Vidal. Hybrid ai approach for counterfactual prediction over knowledge graphs for personal healthcare. In *Workshop on Artificial Intelligence and Data Science for Healthcare: Bridging Data-Centric AI and People-Centric Healthcare. In conjunction with KDD*, 2024.

[11] Fotis Aisopos, Samaneh Jozashoori, Emetis Niazmand, Disha Purohit, Ariam Rivas, Ahmad Sakor, Enrique Iglesias, Dimitrios Vogiatzis, Ernestina Menasalvas, Alejandro Rodríguez González, Guillermo Vigueras, Daniel Gómez-Bravo, Maria Torrente, Roberto Hernández López, Mariano Provencio Pulla, Athanasios Dalianis, Anna Triantafillou, Georgios Paliouras, and Maria-Esther Vidal. Knowledge graphs for enhancing transparency in health data ecosystems. *Semantic Web*, 14(5):943–976, 2023.

[12] Maria-Esther Vidal, Emetis Niazmand, Philipp D. Rohde, Enrique Iglesias, and Ahmad Sakor. *Challenges for Healthcare Data Analytics Over Knowledge Graphs*. 2023.

[13] Payal Chandak, Kexin Huang, and Marinka Zitnik. Building a knowledge graph to enable precision medicine. *bioRxiv*, 2022.

[14] Ahmad Sakor, Samaneh Jozashoori, Emetis Niazmand, Ariam Rivas, Konstantinos Bougiatiotis, Fotis Aisopos, Enrique Iglesias, Philipp D. Rohde, Trupti Padiya, Anastasia Krithara, Georgios Paliouras, and Maria-Esther Vidal. Knowledge4covid-19: A semantic-based approach for constructing a covid-19 related knowledge graph from various sources and analyzing treatments' toxicities. *Journal of Web Semantics*, 75:100760, 2023.

[15] Enrique Iglesias, Samaneh Jozashoori, David Chaves-Fraga, Diego Collarana, and Maria-Esther Vidal. SDM-RDFizer: An RML Interpreter for the Efficient Creation of RDF Knowledge Graphs. In *CIKM '20: Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, New York, USA, 2020. ACM.

[16] Mónica Figuera, Philipp D. Rohde, and Maria-Esther Vidal. Trav-SHACL: Efficiently Validating Networks of SHACL Constraints. In *The Web Conference*, pages 3337–3348, New York, NY, USA, 2021. ACM.

[17] Petar Ristoski and Heiko Paulheim. Semantic web in data mining and knowledge discovery: A comprehensive survey. *J. Web Semant.*, 2016.

[18] Ariam Rivas and Maria-Esther Vidal. Capturing knowledge about drug-drug interactions to enhance treatment effectiveness. In *Proceedings of the 11th Knowledge Capture Conference*, K-CAP '21, New York, NY, USA, 2021. Association for Computing Machinery.

[19] Tiffany J. Callahan, Ignacio J. Tripodi, Adrianne L. Stefanski, Luca Cappelletti, Sanya B. Taneja, Jordan M. Wyrwa, Elena Casiraghi, Nicolas A. Matentzoglu, Justin Reese, Jonathan C. Silverstein, Charles Tapley Hoyt, Richard D. Boyce, Scott A. Malec, Deepak R. Unni, Marcin P. Joachimiak, Peter N. Robinson, Christopher J. Mungall, Emanuele Cavalleri, Tommaso Fontana, Giorgio Valentini, Marco Mesiti, Lucas A. Gillenwater, Brook Santangelo, Nicole A. Vasilevsky, Robert Hoehndorf, Tellen D. Bennett, Patrick B. Ryan, George Hripcsak, Michael G. Kahn, Michael Bada, William A. Baumgartner Jr au2, and Lawrence E. Hunter. An open-source knowledge graph ecosystem for the life sciences, 2024.

[20] Calvo Virginia, Niazmand Emetis, Carcereny Enric, Rodriguez-Abreu Delvys, Cobo Manuel, López-Castro Rafael, Guirado María, Camps Carlos, Laura Ortega Ana, Bernabé Reyes, Massutí Bartomeu, Garcia-Campelo Rosario, del Barco Edel, Luis González-Larriba José, Bosch-Barrera Joaquim, Martínez Marta, Torrente María, Vidal María-Esther, and Provencio Mariano. Family history of cancer and lung cancer: Utility of

big data and artificial intelligence for exploring the role of genetic risk. *Lung Cancer*, 2024.

[21] Adrianna Janik, Maria Torrente, Luca Costabello, Virginia Calvo, Brian Walsh, Carlos Camps, Sameh K. Mohamed, Ana L. Ortega, Vít Nováček, Bartomeu Massutí, Pasquale Minervini, M. Rosario Garcia Campelo, Edel del Barco, Joaquim Bosch-Barrera, Ernestina Menasalvas, Mohan Timilsina, and Mariano Provencio. Machine learning–assisted recurrence prediction for patients with early-stage non–small-cell lung cancer. *JCO Clinical Cancer Informatics*, (7):e2200062, 2023.

[22] Liyan Pan, Guangjian Liu, Fangqin Lin, Shuling Zhong, Huimin Xia, Xin Sun, and Huiying Liang. Machine learning applications for prediction of relapse in childhood acute lymphoblastic leukemia. *Scientific Reports*, 7, 2017.

[23] Yang Yang, Li Xu, Liangdong Sun, Peng Zhang, and Suzanne S. Farid. Machine learning application in personalised lung cancer recurrence and survivability prediction. *Computational and Structural Biotechnology Journal*, 20:1811–1820, 2022.

[24] Satya Maurya, Pushpendra Sisodia, Rahul Mishra, and Devesh singh. Performance of machine learning algorithms for lung cancer prediction: a comparative approach. *Scientific Reports*, 14, 08 2024.

[25] Yawei Li, Xin Wu, Ping Yang, Guoqian Jiang, and Yuan Luo. Machine learning applications in lung cancer diagnosis, treatment and prognosis, 2022.

[26] Leo Breiman. Random forests. *Machine learning*, 45, 2001.

[27] Akhilesh Vyas, Fotis Aisopos, Maria-Esther Vidal, Peter Garrard, and Georgios Paliouras. Identifying the presence and severity of dementia by applying interpretable machine learning techniques on structured clinical records. *BMC Medical Informatics Decis. Mak.*, 22(1):271, 2022.

[28] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2016. ACM.

[29] Ashley Suh, Gabriel Appleby, Erik W. Anderson, Luca Finelli, Remco Chang, and Dylan Cashman. Are metrics enough? guidelines for communicating and visualizing predictive models to subject matter experts. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–16, 2023.

[30] Harry X. Li, Gabriel Appleby, Camelia Daniela Brumar, Remco Chang, and Ashley Suh. Knowledge Graphs in Practice: Characterizing their Users, Challenges, and Visualization Opportunities. *IEEE Trans. Vis. Comput. Graph.*, 30(1):584–594, 2024.

[31] Disha Purohit and Maria-Esther Vidal. Mining symbolic rules to explain lung cancer treatments. In Catia Pesquita, Hala Skaf-Molli, Vasilis Efthymiou, Sabrina Kirrane, Axel Ngonga, Diego Collarana, Renato Cerqueira, Mehwish Alam, Cássia Trojahn, and Sven Hertling, editors, *The Semantic Web: ESWC 2023 Satellite Events - Hersonissos, Crete, Greece, May 28 - June 1, 2023, Proceedings*, volume 13998 of *Lecture Notes in Computer Science*, pages 69–74. Springer, 2023.

[32] Ariam Rivas, Diego Collarana, Maria Torrente, and Maria-Esther Vidal. A neuro-symbolic system over knowledge graphs for link prediction. *Semantic Web Journal. Special Issue on Neuro-Symbolic Artificial Intelligence and the Semantic Web*, pages 1–25, 2023.

[33] Yashrajsinh Chudasama, Disha Purohit, Philipp D. Rohde, Julian Gercke, and Maria-Esther Vidal. InterpretME: A Tool for Interpretations of Machine Learning Models Over Knowledge Graphs. *Semantic Web Journal. Special Issue on Tools & Systems*, 2024.

[34] Yashrajsinh Chudasama, Disha Purohit, Philipp D. Rohde, and Maria-Esther Vidal. Enhancing interpretability of machine learning models over knowledge graphs. In Neha Keshan, Sebastian Neumaier, Anna Lisa Gentile, and Sahar Vahdati, editors, *Proceedings of the Posters and Demo Track of the 19th International Conference on Semantic Systems co-located with 19th International Conference on Semantic Systems (SEMANTiCS 2023), Leipzig, Germany, September 20 to 22, 2023*, volume 3526 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2023.

[35] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[36] Yashrajsinh Chudasama, Disha Purohit, Philipp D. Rohde, Enrique Iglesias, Maria Torrente, and Maria-Esther Vidal. Semantically describing predictive models for interpretable insights into lung cancer relapse. In Angelo A. Salatino, Mehwish Alam, Femke Ongenae, Sahar Vahdati, Anna Lisa Gentile, Tassilo Pellegrini, and Shufan Jiang, editors, *Knowledge Graphs in the Age of Language Models and Neuro-Symbolic AI - Proceedings of the 20th International Conference on Semantic Systems, 17-19 September 2024, Amsterdam, The Netherlands*, volume 60 of *Studies on the Semantic Web*, pages 142–158. IOS Press, 2024.

[37] M. Badawy, N. Ramadan, and H.A. Hefny. Healthcare predictive analytics using machine learning and deep learning techniques: a survey. *Journal of Electrical Systems and Information Technology*, 2023.

[38] J. Bajwa, U. Munir, A. Nori, et al. Artificial intelligence in healthcare: transforming the practice of medicine. *Future healthcare journal*, 2021.

[39] T. Davenport and R. Kalakota. The potential for artificial intelligence in healthcare. *Future healthcare journal*, 2019.

[40] J. Pearl. *Causality*. Cambridge university press, 2009.

[41] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

[42] Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

[43] Babak Salimi, Harsh Parikh, Moe Kayali, Lise Getoor, Sudeepa Roy, and Dan Suciu. Causal relational learning. In *Proceedings of the 2020 ACM SIGMOD international conference on management of data*, pages 241–256, 2020.

[44] Hao Huang. Causal relationship over knowledge graphs. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 5116–5119, 2022.

[45] H. Huang and M.E. Vidal. Causekg: A framework enhancing causal inference with implicit knowledge deduced from knowledge graphs. *IEEE Access*, 2024.

[46] D.M. Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 2002.

[47] Barbra A Dickerman, Issa J Dahabreh, Krystal V Cantos, Roger W Logan, Sara Lodi, Christopher T Rentsch, Amy C Justice, and Miguel A Hernán. Predicting counterfactual risks under hypothetical treatment strategies: an application to hiv. *European journal of epidemiology*, 37(4):367–376, 2022.

[48] Qing Gao, Luyu Yang, Mingjun Lu, Renjing Jin, Huan Ye, and Teng Ma. The artificial intelligence and machine learning in lung cancer immunotherapy. *Journal of Hematology and Oncology*, 16(1), May 2023.

[49] Line Farah, Juliette M. Murris, Isabelle Borget, Agathe Guilloux, Nicolas M. Martelli, and Sandrine I.M. Katsahian. Assessment of performance, interpretability, and explainability in artificial intelligence–based health technologies: What healthcare stakeholders need to know. *Mayo Clinic Proceedings: Digital Health*, 1(2):120–138, 2023.

[50] Kerstin Lenhof, Lea Eckhart, Lisa-Marie Rolli, Andrea Volkamer, and Hans-Peter Lenhof. Reliable anti-cancer drug sensitivity prediction and prioritization. *Scientific Reports*, 14(1), May 2024.

[51] Shuroug A. Alowais, Sahar S. Alghamdi, Nada Alsuhebany, Tariq Alqahtani, Abdulrahman I. Alshaya, Sumaya N. Almohareb, Atheer Aldairem, Mohammed Alrashed, Khalid Bin Saleh, Hisham A. Badreldin, Majed S. Al Yami, Shmeylan Al Harbi, and Abdulkareem M. Albekairy. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC Medical Education*, 23(1), September 2023.

[52] Marzyeh Ghassemi, Tristan Naumann, Peter Schulam, Andrew L Beam, Irene Y Chen, and Rajesh Ranganath. Practical guidance on artificial intelligence for health-care data. *The Lancet Digital Health*, 1(4):e157–e159, August 2019.

[53] Ilaria Tiddi and Stefan Schlobach. Knowledge graphs as tools for explainable machine learning: A survey. *Artificial Intelligence*, 302:103627, January 2022.

[54] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. *Knowledge Graphs*. Synthesis Lectures on Data, Semantics, and Knowledge. Morgan & Claypool Publishers, 2021.

[55] Mauro Giuffrè and Dennis L. Shung. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *npj Digital Medicine*, 6(1), October 2023.

[56] Allard Oelen, Mohamad Yaser Jaradeh, and Sören Auer. ORKG ASK: a neuro-symbolic scholarly search and exploration system. In *Joint Proceedings of Posters, Demos, Workshops, and Tutorials of the 20th International Conference on Semantic Systems co-located with 20th International Conference on Semantic Systems (SEMANTiCS 2024), Amsterdam, The Netherlands, September 17-19, 2024*, volume 3759 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2024.

[57] Michael van Bekkum, Maaike de Boer, Frank van Harmelen, André Meyer-Vitali, and Annette ten Teije. Modular design patterns for hybrid learning and reasoning systems: a taxonomy, patterns and use cases. *CoRR*, 2021.

[58] Disha Purohit, Yashrajsinh Chudasama, Ariam Rivas, and Maria-Esther Vidal. Sparkle: Symbolic capturing of knowledge for knowledge graph enrichment with learning. In *Proceedings of the 12th Knowledge Capture Conference 2023*, K-CAP '23, page 44–52, New York, NY, USA, 2023. Association for Computing Machinery.

[59] Mehdi Ali, Max Berrendorf, Charles Tapley Hoyt, Laurent Vermue, Sahand Sharifzadeh, Volker Tresp, and Jens Lehmann. PyKEEN 1.0: A Python Library for Training and Evaluating Knowledge Graph Embeddings. *Journal of Machine Learning Research*, 2021.

[60] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Durán, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. NIPS'13, page 2787–2795, Red Hook, NY, USA, 2013. Curran Associates Inc.

[61] Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. Knowledge graph embedding via dynamic mapping matrix. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers)*, pages 687–696, 2015.

[62] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. Knowledge graph embedding by translating on hyperplanes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 28(1), 2014.

[63] Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. Rotate: Knowledge graph embedding by relational rotation in complex space, 2019.

[64] Yashrajsinh Chudasama. Exploiting semantics for explaining link prediction over knowledge graphs. In Catia Pesquita, Hala Skaf-Molli, Vasilis Efthymiou, Sabrina Kirrane, Axel Ngonga, Diego Collarana, Renato Cerqueira, Mehwish Alam, Cássia Trojahn, and Sven Hertling, editors, *The Semantic Web: ESWC 2023 Satellite Events - Hersonissos, Crete, Greece, May 28 - June 1, 2023, Proceedings*, volume 13998 of *Lecture Notes in Computer Science*, pages 321–330. Springer, 2023.

[65] Farahnaz Akrami, Mohammed Samiul Saeef, Qingheng Zhang, Wei Hu, and Chengkai Li. Realistic re-evaluation of knowledge graph completion methods: An experimental study. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, SIGMOD '20, page 1995–2010, New York, NY, USA, 2020. Association for Computing Machinery.

[66] Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4(none), January 2010.

[67] Jonathan Lajus, Luis Galárraga, and Fabian Suchanek. Fast and Exact Rule Mining with AMIE 3. In *The Semantic Web*, 2020.

[68] Stefan Decker, Prasenjit Mitra, and Sergey Melnik. Framework for the semantic web: an rdf tutorial. *IEEE Internet Computing*, 4(6):68–73, 2000.

[69] Boris Motik, Peter F Patel-Schneider, Bijan Parsia, Conrad Bock, Achille Fokoue, Peter Haase, Rinke Hoekstra, Ian Horrocks, Alan Ruttenberg, Uli Sattler, et al. Owl 2 web ontology language: Structural specification and functional-style syntax. *W3C recommendation*, 27(65):159, 2009.

[70] Pierre-Yves Vandenbussche, Ghislain A Atemezing, María Poveda-Villalón, and Bernard Vatant. Linked open vocabularies (lov): a gateway to reusable semantic vocabularies on the web. *Semantic Web*, 8(3):437–452, 2017.

[71] Sondos Mahmoud Bsharat, Aidar Myrzakhan, and Zhiqiang Shen. Principled instructions are all you need for questioning llama-1/2, gpt-3.5/4. *arXiv preprint arXiv:2312.16171*, 2023.

[72] P. Spirtes, C. Glymour, and R. Scheines. *Causation, prediction, and search*. 2001.

[73] P. Hoyer, D. Janzing, J.M. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 2008.

[74] Michael Levandowsky and David Winter. Distance between sets. *Nature*, 234(5323):34–35, 1971.

[75] Peter Flach and Meelis Kull. Precision-recall-gain curves: Pr analysis done right. *Advances in neural information processing systems*, 28, 2015.

[76] Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4, 2009.

[77] OpenAI. Chatgpt, 2024. Accessed: 2024-05.

[78] Ankur Ankan and Abinash Panda. pgmpy: Probabilistic graphical models using python. In *SciPy*, pages 6–11. Citeseer, 2015.

[79] In Jae Myung. Tutorial on maximum likelihood estimation. *Journal of mathematical Psychology*, 47(1):90–100, 2003.

[80] D.M. Hawkins. The problem of overfitting. *Journal of chemical information and computer sciences*, 2004.

YASHARAJSINH CHUDASAMA Email: yashrajsinh.chudasama@tib.eu

M.Sc. Yasharajsinh Chudasama is a research associate at the Scientific Data Management (SDM) group at the TIB-Leibniz Information Centre for Science and Technology. He is also doing his doctoral studies at Leibniz University of Hannover under the supervision of Prof. Dr. Maria-Esther Vidal. Yashrajsinh holds a master's degree (M.Sc.) Mechatronics from Leibniz University Hannover (LUH) and a bachelor's degree (B.E.) Mechatronics from Gujarat, India. He researches data management, sub-symbolic learning, explainability, and neuro-symbolic AI over knowledge graphs. He is working on a TrustKG project funded by the Leibniz Association.

HAO HUANG Email: hao.huang@tib.eu

Hao Huang is a research assistant at the Scientific Data Management (SDM) group of TIB – Leibniz Information Centre for Science and Technology and a doctoral candidate at the Faculty of Electrical Engineering and Computer Science of Leibniz Universität Hannover (LUH), Germany. He has earned a master degree in Computer Science and Technology from the South China University of Technology (SCUT), China, and another master degree in Data Science from Université de Nantes, France. His undergraduate studies culminated in a Bachelor of Science degree in Information Security from Hunan University of Science and Technology. Hao has previously engaged in research on Ontology Learning and Topic Modeling. Currently, Hao is concentrating on the interdisciplinary area of causal inference and Knowledge Graphs (KGs), with a particular interest in combining the semantics of KGs with causal inference techniques to enhance the trustability and interpretability in answering causal questions over KGs.

DISHA PUROHIT Email: disha.purohit@tib.eu

M.Sc. Disha Purohit is a research associate at the Scientific Data Management (SDM) group at TIB-Leibniz Information Centre for Science and Technology. She is also doing her doctoral studies at Leibniz University Hannover under the supervision of Prof. Dr. Maria-Esther Vidal. Disha Purohit holds a master's degree (M.Sc.) Internet Technologies and Information Systems (ITIS) from Leibniz University Hannover (LUH) and a bachelor's degree (B.E.) Computer Engineering from Mumbai, Maharashtra, India. She researches data management, inductive learning, specifically symbolic learning, and discovering causal patterns from the knowledge graphs. She is working on a TrustKG (a project funded by the Leibniz Association) project and the P4-LUCAT project funded by ERAMed.

PROF. DR. MARIA-ESTHER VIDAL **Email: maria.vidal@tib.eu**

Prof. Dr. Maria-Esther Vidal is a full professor at the Leibniz University Hannover and leads the Scientific Data Management (SDM) group at TIB-Leibniz Information Centre for Science and Technology. She is also a member of the L3S Research Centre and a full professor (retired) at Universidad Simón Bolívar (USB), Venezuela. She researches data management, semantic data integration, and machine learning over knowledge graphs. Maria-Esther is a co-author of more than 240 peer-reviewed articles in Semantic Web, Databases, and Artificial Intelligence. She has been awarded the Science Award on Responsible Research by Stifterverband with the recommendation of the Leibniz Association and the program "Leibniz Best Minds: Programme for Women Professors" supported by the Leibniz Association, Germany. Maria-Esther is also actively shaping her research communities. She has been an editorial board member of renowned journals (e.g., JWS, JDIQ) and general chair, co-chair, and senior reviewer of major scientific events (e.g., ESWC, WWW, ISWC, and AAAI). Under her direction, her team has developed technologies of predominant relevance in the whole process of knowledge graph creation from heterogeneous data and query processing. She serves as an expert in several advisory boards, summer schools, and doctoral consortiums. She has advised more than 28 doctoral students and more than 120 Master's and bachelor's students in Computer Science. She has been a doctoral committee member in France, Italy, Sweden, Spain, the Netherlands, Germany, Ireland, Argentina, Uruguay, and Venezuela.