## Blog post Universidad Politécnica de Madrid (UPM) – April 2024

# Exploring Knowledge Graphs: A Primer on Representing Lung Cancer and Environmental Factors


In the era of big data, knowledge graphs have emerged as powerful tools for organizing, integrating, and deriving insights from vast amounts of information. Particularly in the medical domain, where the complexity of diseases like lung cancer demands comprehensive understanding, knowledge graphs offer a promising approach. This article serves as an introduction to knowledge graphs and their application in representing medical concepts related to lung cancer, including factors influencing its development.

At its core, a knowledge graph is a structured representation of knowledge that encapsulates entities, their attributes, and the relationships between them. Unlike traditional databases, knowledge graphs are designed to capture semantic connections, allowing for richer and more nuanced insights. Entities in a knowledge graph can range from simple concepts like genes or proteins to complex entities such as diseases or environmental factors.

In the context of LUCIA project, in which lung cancer is studied to find new risk factors, a knowledge graph can incorporate various entities relevant to the disease. This includes molecular factors such as specific gene mutations or protein biomarkers, data extracted from electronic health records (EHR) of patients like smoking habits, tumor stage or histological subtype, as well as environmental influences such as air pollution levels. By linking these entities through meaningful relationships, a knowledge graph provides a holistic view of the disease landscape, enabling researchers to explore intricate connections and identify potential targets for intervention.

Although it is well known that smoking stands out as the main cause of lung cancer, other factors such as the presence of radon and other pollutants in the air and exposure to asbestos have also been implicated in increasing the risk of lung cancer. Understanding the interaction between these factors and genetic predispositions is crucial to elucidate the underlying mechanisms of the disease and to design effective prevention strategies. The integration of these data, together with the geographical location, can provide context for different aspects that occur at a moment and over an extended duration.

## Application of Knowledge Graphs in Lung Cancer Research

Knowledge graphs offer several advantages in the study of lung cancer and its associated environmental factors. Firstly, they facilitate data integration from diverse sources, including clinical databases, genomic repositories, and environmental monitoring records. By harmonizing heterogeneous data into a unified framework, knowledge graphs enable researchers to uncover hidden patterns and correlations that may not be apparent through isolated analyses.

Secondly, knowledge graphs support semantic querying and inference, allowing researchers to pose complex questions and derive actionable insights. For instance, one could query the graph to identify genes associated with both tobacco exposure and lung cancer susceptibility, thereby revealing potential molecular pathways underlying the disease. Similarly, researchers can explore the impact of environmental pollutants on gene expression patterns or clinical outcomes, shedding light on their contribution to disease progression. Despite their potential, knowledge graphs face several challenges in the context of lung cancer research. Integrating heterogeneous data sources while ensuring data quality and interoperability remains a challenging task. Thus, understanding the importance of registrations for lung cancer's incidence, prevalence, and mortality offers vital insights for healthcare, policy, and research, helping to understand the disease burden and guide resource allocation and public health efforts to reduce its impact. Looking ahead, advancements in data standardization, ontological modelling, and artificial intelligence will fuel the development of more sophisticated knowledge graphs tailored to the complexities of lung cancer. By harnessing the collective wisdom encoded in these graphs, researchers can accelerate discoveries, improve patient outcomes, and ultimately combat this devastating disease.

As part of the LUCIA project, the Universidad Politécnica de Madrid (UPM) is currently engaged in developing the structure of a knowledge graph aimed at modeling data pertinent to lung cancer and its associated risk factors. Concurrently, efforts are underway to establish a repository containing public data sets, which will serve as foundational resources for populating the knowledge graph. Looking ahead, the roadmap involves leveraging advanced machine learning algorithms to autonomously uncover novel connections within the graph, potentially unveiling hidden risk factors associated with lung cancer that have yet to be elucidated.

The LUCIA project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101096473. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the Health and Digital Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.