A Dynamic Knowledge Graph of Interaction Between Pollution and Cardiovascular Diseases

Sudeshna Jana[1], **Anunak Roy** [2], **Manjira Sinha** [1], **Tirthankar Dasgupta** [1],
[1]TCS Research, [2] IIT Kharagpur,
Correspondence: { sudeshna.jana, sinha.manjira, dasgupta.tirthankar} @tcs.com, anunakroy@kgpian.iitkgp.ac.in

Abstract

In recent decades, environmental pollution has become a pressing global health concern. According to the World Health Organization (WHO), a significant portion of the population is exposed to air pollutant levels exceeding safety guidelines. Cardiovascular diseases (CVDs) — including coronary artery disease, heart attacks, and strokes — are particularly significant health effects of this exposure. In this paper, we investigate the effects of air pollution on cardiovascular health by constructing a dynamic knowledge graph based on extensive biomedical literature. This paper provides a comprehensive exploration of entity identification and relation extraction, leveraging advanced language models. Additionally, we demonstrate how in-context learning with large language models can enhance the accuracy and efficiency of the extraction process. The constructed knowledge graph enables us to analyze the relationships between pollutants and cardiovascular diseases over the years, providing deeper insights into the long-term impact of cumulative exposure, underlying causal mechanisms, vulnerable populations, and the role of emerging contaminants in worsening various cardiac outcomes.

1 Introduction

Over the past few decades, the increasing levels of environmental pollution have emerged as a formidable global health crisis. The World Health Organization (WHO) has reported that nearly 99% of the world's population breathes air that contains pollutant levels exceeding established safety guidelines (Organization et al., 2016). This exposure is particularly severe in low- and middle-income countries, where industrial activities, urbanization, and insufficient regulatory measures exacerbate air quality issues. Among the various health effects attributed to pollution, its impact on cardiovascular diseases (CVDs) is particularly concerning (Rajagopalan et al., 2018). CVDs encompass a wide range of conditions, including coronary artery disease, heart attacks, strokes, and heart failure. These diseases are not only prevalent but also represent a leading cause of morbidity and mortality worldwide. The mechanisms by which air pollutants influence cardiovascular health are multifaceted, involving both direct effects on the cardiovascular system and indirect effects through exacerbating existing risk factors such as hypertension and diabetes.

Recent studies have demonstrated that pollutants such as fine particulate matter (PM), heavy metals, and toxic gases have a significant impact on cardiac health (Basith et al., 2022; Zhang et al., 2022, 2016). Research conducted by the United States Environmental Protection Agency (EPA) indicates that exposure to elevated concentrations of $PM_{2.5}$, even over a short duration of a few hours to weeks, can trigger heart attacks and fatalities associated with cardiovascular diseases. Prolonged exposure to these pollutants is linked to an increased risk of cardiovascular mortality and a reduction in life expectancy (Beelen et al., 2014). Furthermore, a substantial body of epidemiological evidence reveals a strong correlation between air pollutants and rising rates of cardiovascular diseases, including heart failure (Jia et al., 2023). Animal studies also corroborate these findings, illustrating that exposure to pollutants can elevate the likelihood of conditions such as thrombosis and atherosclerosis (Sun et al., 2005). These insights underscore the urgent need for comprehensive public health interventions aimed at reducing air pollution and mitigating its adverse effects on cardiovascular health.

However, current findings do not demonstrate how cumulative exposure impacts cardiovascular health over decades, the precise causal pathways by which these pollutants lead to cardiovascular damage. Additionally, it is crucial to identify which populations are most vulnerable to specific pollutants and the reasons behind their susceptibil-

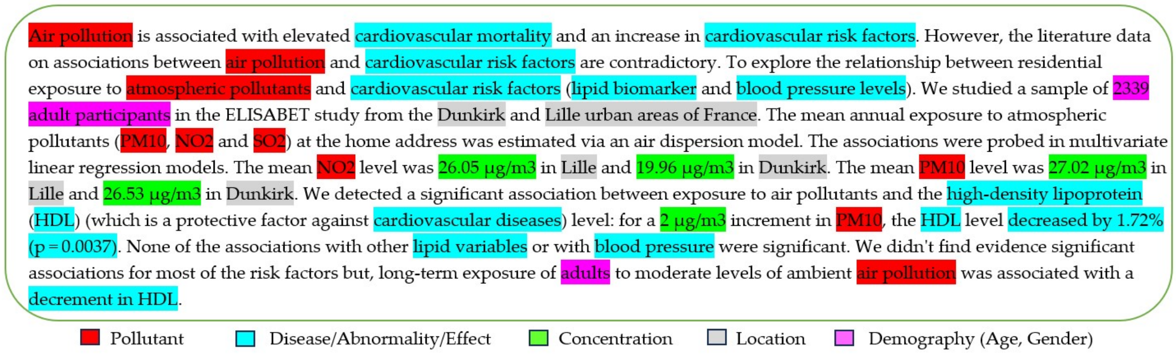


Figure 1: Sample PubMed abstract with highlighted entities.

ity, which may include genetic predispositions, environmental factors, pre-existing health conditions, and lifestyle choices. Moreover, there is a growing concern regarding new pollutants and emerging contaminants—such as microplastics and various industrial chemicals—that remain under-researched in the context of cardiovascular health. Motivated by these gaps, In this study, we seek to address several of these issues by constructing a comprehensive knowledge graph ‘PollCardioKG’ that visualizes the causal relationships between pollutants and cardiovascular diseases, based on an extensive review of scientific PubMed abstracts. We also investigate how varying concentrations of pollutants affect cardiovascular health differently across diverse populations. This knowledge will enhance our understanding of the mechanisms by which pollutants contribute to cardiovascular diseases and inform the development of more effective public health strategies to mitigate their impact.

The remainder of this article is organized as follows: In the next section, we provide a brief overview of related works conducted by previous researchers in this field. In Section 3, we detail our methodology for dataset collection, data representation, and knowledge graph generation. Section 4 presents our results along with a thorough analysis of the experimental findings. Finally, in Section 5, we draw conclusions based on our analysis, and in Section 6, we discuss the limitations of our study.

2 Related works

2.1 Empirical Experiments on the effect of pollutants on cardiovascular diseases

Over the years, a wealth of empirical studies conducted by various researchers has provided critical insights into the relationship between pollutants and cardiovascular diseases. For instance, Feng and Yang (Feng and Yang, 2012) demonstrated a clear association between exposure to fine particulate matter ($PM_{2.5}$ and PM_{10}) and an increased risk of developing cardiovascular diseases. In a subsequent study, Miller et al. (Miller and Newby, 2020) found that exposure to particulate matter, particularly ultrafine particles originating from diesel emissions, is linked to heightened cardiovascular risks, including coronary artery disease, hypertension, and arrhythmias. More recently, Lederer et al. (Lederer et al., 2021) emphasized the significant role of pollutants such as $PM_{2.5}$ and nitrogen dioxide (NO_2) in elevating cardiovascular morbidity and mortality rates. Their findings indicated that $PM_{2.5}$ is associated with oxidative stress and thrombosis, while NO_2 is linked to systemic inflammation and autonomic dysfunction.

2.2 Computational models on the effect of pollutants on cardiovascular diseases

In this era of artificial intelligence, a growing number of researchers are employing various machine learning and deep learning techniques to understand and predict the impacts of pollutants on cardiovascular health. For example, in 2017, Park et al. (Park et al., 2017) developed an Environmental Risk Score (ERS) to estimate the health effects of pollutant mixtures. Their study highlighted the cumulative risks associated with exposure to pollutants such as cadmium, arsenic compounds, cobalt, and barium, particularly in relation to gamma-glutamyl transferase levels and hypertension. More recently, in 2022, Lee et al. (Lee et al., 2022) utilized a range of machine learning models to predict non-accidental mortality due to cardiovascular and respiratory causes, leveraging meteorological data alongside air pollution metrics.

USER: You have been provided with a PubMed Abstract and following instructions. Write a response that appropriately complete the instructions.

Instructions:
You need to perform the following tasks:
1. Identify and list all POLLUTANT NAMES with POLLUTANT LEVEL mentioned in the PubMed abstract.
2. Identify and list all DISEASE NAMES mentioned in the abstract.
3. Capture demographical information – STUDY LOCATION, AGE GROUP and GENDER of the subjects, if available in the abstract.
4. Format the Output: Return in a structured JSON format with the keys as "Pollutant names", "Disease Names", "Study Location", "Population age group" and "Population gender". Put 'N/A' for any missing values.

PubMed Abstract:
Now here is the PubMed abstract:
{abstract}

Figure 2: Prompt template used for entity identification and relation extraction.

While most of these studies have concentrated on the effects of air pollution on the cardiovascular system using location-specific pollution data, our study adopts a more comprehensive approach. By examining the multifaceted impacts of pollutants on cardiovascular diseases across diverse locations and populations, we aim to provide a broader understanding of this critical issue. Our analysis of a wide range of pollution-related scientific literature not only deepens our insights into the established effects of traditional pollutants but also highlights the potential risks associated with new and emerging contaminants. This holistic perspective is essential for developing effective public health strategies to mitigate the adverse impacts of pollution on cardiovascular health.

3 Proposed Methodology

3.1 Downloading PubMed Abstracts

In our work, we have collected a total of 4,716 PubMed abstracts [1] published between 2012 and the present, using the *Metapub* library, focusing on the keywords "cardiovascular disease" and "air pollution". Each abstract is assigned a unique identifier, referred to as 'pmid'. Additionally, we extracted the publication year of each study, enabling us to perform a temporal analysis to track how the relationship between pollutants and cardiovascular health has evolved over time. These abstracts provide a wealth of information, often containing valuable details such as pollutant names, pollutant concentrations, associated cardiovascular diseases, geographical locations, and the specific target populations studied. Figure 1 presents a sample PubMed abstract with various portions of text color-coded to highlight these different entities, offering insight into the richness of the data and the variety of factors considered in these studies.

3.2 Entity Identification

From each of the collected PubMed abstracts, we extracted the following entities and relations:

- **Pollutant Names (P)**: The pollutants mentioned in the abstract.
- **Pollutant Levels (C)**: The concentration levels of the pollutants.
- **Disease Names (D)**: The diseases or health conditions linked to the pollutants.
- **Study Location (Loc)**: The location where the study was conducted.
- **Population Details (S)**: The demographic details of the affected population.

We experimented with various large language models (LLMs) such as —Llama2-7B (Touvron et al., 2023), Llama-3-8B (Dubey et al., 2024), Asclepius-7B (Wang et al., 2024), PMC-LLaMA-7B (Wu et al., 2024), GPT-4 (Waisberg et al., 2023) —for this entity extraction task. A variety of prompts were tested, and the one displayed in Figure 2 yielded the best results, as measured by human evaluation.

[1] https://pubmed.ncbi.nlm.nih.gov/

3.3 Entity Standardization

The pollutants and diseases extracted from the LLM are not always in a standardized form. For example, the pollutant “ultrafine particles” may be referred as “ufp”, “pm $< 0.01\mu m$ aerodynamic diameter” in different articles. Similarly, disease names like “Hypertension” may appear as “High Blood Pressure”, “Arterial Hypertension”, or “Hypertensive Disorder” in various research.

To address this variability, all extracted entities were standardized using the Unified Medical Language System (UMLS) Metathesaurus (Schuyler et al., 1993), which provides a comprehensive list of terms and assigns a Concept Unique Identifier (CUI) to each. However, we found that some entities did not match any UMLS concept exactly. To resolve this, we applied an approximate string-matching algorithm based on the Levenshtein distance measure (Yujian and Bo, 2007) to identify the closest UMLS concept. For entities that could not be mapped to any UMLS concept, we created unique identifiers to ensure that no entities were overlooked.

3.4 Relation Extraction

Subsequently, we analyze the relationships between all extracted pollutant-disease pairs, classifying them into categories such as ‘cause’, ‘may cause’, and ‘does not cause’. In this study, we employ a Chain of Thought (CoT) framework with In-Context Learning (ICL) to leverage the inherent capabilities of LLMs for relation extraction within a few-shot learning paradigm. This approach strategically incorporates in-context learning by embedding a set of 10 curated examples into the model’s prompt. Each example consists of a text passage, a detailed step-by-step explanation of the reasoning process, and the resulting relation triples, which guide the model in identifying and reasoning about the relationships among entities in the given texts. The prompt examples cover three primary relation types—‘cause’, ‘may cause’, and ‘does not cause’—and involve entities identified as Pollutant Names (P) and Disease Names (D).

For illustration, consider the following example included in the model’s prompt:

Instruction: Identify the relationship among the entities [Pollutant: NO2 and Disease: ischemic stroke] in the given text, categorizing it as one of cause’, may cause’, or ‘does not cause’, and provide a reasonable explanation.

Text: "Polish smog is a specific type of air pollution present in Eastern Poland, which causes particularly adverse cardiovascular effects. Additionally, PM and nitrogen dioxide (NO2) have an impact on mortality due to acute coronary syndrome (ACS) and ischemic stroke (IS)."

Explanation: PM and nitrogen dioxide (NO2) impact mortality due to acute coronary syndrome (ACS) and ischemic stroke (IS).

Relations: [[“NO2”, “causes”, “ischemic stroke”]]

This integration of in-context learning and logical reasoning underscores the robust capabilities of advanced language models in addressing intricate challenges in natural language understanding.

3.5 Building Knowledge Graph

Once all pollutant and disease names were standardized, we constructed a knowledge graph to provide a structured representation of the complex relationships between these entities (Fensel et al., 2020). In this graph, pollutants and diseases are represented as nodes, while the relationships between them (e.g., causes or may cause) are depicted as directed edges from pollutants to diseases. Where available, additional attributes such as pollutant concentration levels and demographic information (e.g., location, age group, gender) were attached to the corresponding nodes. This knowledge graph facilitates a clearer understanding of how specific pollutants are linked to diseases and enables exploration of broader patterns and connections. It also highlights how certain diseases may be related through common pollutants and identifies key pollutants and diseases, thereby revealing the most significant relationships within the data.

4 Results and Discussions

In this section, we evaluate the performance of various large language models (LLMs) in extracting entities and their relationships. Furthermore, we conducted a temporal analysis to examine the variation of each pollutant, disease, and their interrelationships over time.

After the pre-processing step outlined in Section 3, which involved 4,716 PubMed abstracts, we compiled a comprehensive list of 300 unique pollutants and approximately 936 unique disease names. We then evaluated the performance of various large language models (LLMs) for entity extraction, comparing their outputs with expert anno-

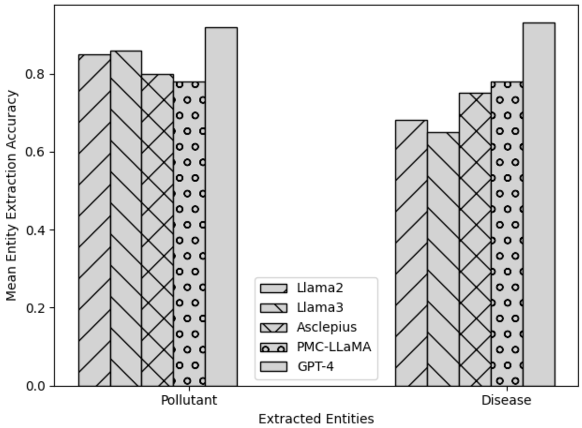


Figure 3: Comparison of Mean Entity Extraction Accuracy across different LLMs.

tations. Our analysis revealed that both LLaMa-2-7B and LLaMa-3-8B models often struggled with accurately identifying disease names, frequently mislabeling conditions. For example, the LLaMa models generated 'Stroke' in place of 'Decreased HDL level' and identified 'cardiovascular risk' instead of 'ischemic heart disease'.

To quantify the performance of these models, we calculated the Mean Entity Extraction Accuracy (MEEA), defined by $\frac{1}{N}\sum_{i=1}^{N}\frac{n(\hat{E}_i)}{n(E_i)}$, where N represents the total number of abstracts, $n(E)$ denotes the number of entities annotated by human experts, and $n(\hat{E})$ represents the number of entities correctly identified by the LLMs. Among the evaluated models, GPT-4 emerged as the most effective in accurately identifying both pollutant and disease names, as illustrated in Figure 3.

Subsequently, we constructed the knowledge graph, named 'PollCardioKG', which encapsulates a total of 1,907 nodes, consisting of 1,029 pollutant nodes and 878 disease nodes, connected by 5,618 edges. This graph provides a comprehensive view of the complex relationships between pollutants and cardiovascular diseases. To enhance its granularity, multiple nodes were created for the same pollutant to represent varying concentration levels, where such data was available. This structure allows for the differentiation of effects based on pollutant exposure levels, offering deeper insights into potential dose-response relationships.

Figure 4 presents a snippet of the knowledge graph, where pollutant nodes are highlighted in red, and disease nodes are marked in blue. The directed edges between these nodes represent relationships such as causality or association.

Furthermore, we utilized the study year of each abstract as the time variable for this analysis to examine how pollutant levels and disease rates have changed over time. This approach allows us to identify long-term trends and patterns, as well as to detect any unusual or unexpected fluctuations in both pollutant concentrations and disease incidence.

4.1 Longitudinal Analysis of Pollutants

Figure 5(a) illustrates the trends in the occurrence of the top five pollutants identified in the PubMed abstracts from 2012 to 2024. The data indicate a clear upward trajectory for $PM_{2.5}$ exposure, which peaked in 2022 before experiencing a slight decline in subsequent years. This suggests that while $PM_{2.5}$ remains a significant concern, there may be emerging factors influencing its concentration or reporting. Nitrogen Dioxide NO_2) has similarly demonstrated a steady increase, particularly accelerating after 2020, highlighting potential shifts in air quality and pollution sources during this period. In contrast, pollutants such as ozone (O_3), carbon monoxide (CO), and sulfur dioxide (SO_2) have exhibited relatively stable trends, characterized by minor fluctuations.

Furthermore, Figure 5(b) depicts the trends of previously less significant pollutants that have gained increased attention in recent years. Notably, PM_1 has experienced a dramatic rise in mentions starting from 2019, culminating in a peak in 2023 and maintaining elevated levels into 2024. This surge in interest may reflect growing awareness of the health impacts associated with ultrafine particles. Additionally, ammonium and cadmium have shown significant increases in their occurrences within the discussions during this timeframe. Ammonium levels began to escalate after 2016, suggesting a potential link to agricultural practices or changes in atmospheric chemistry. Cadmium, in particular, has seen a remarkable rise in mentions from 2020 to 2024, signaling a need for further investigation into its sources and health implications. Overall, these trends underscore the dynamic nature of pollutant occurrences and highlight the importance of ongoing research and monitoring efforts to address emerging environmental health concerns.

4.2 Longitudinal Analysis of Diseases

Figure 6(a) highlights trends of the top five diseases in PubMed abstracts from 2012 to 2024,

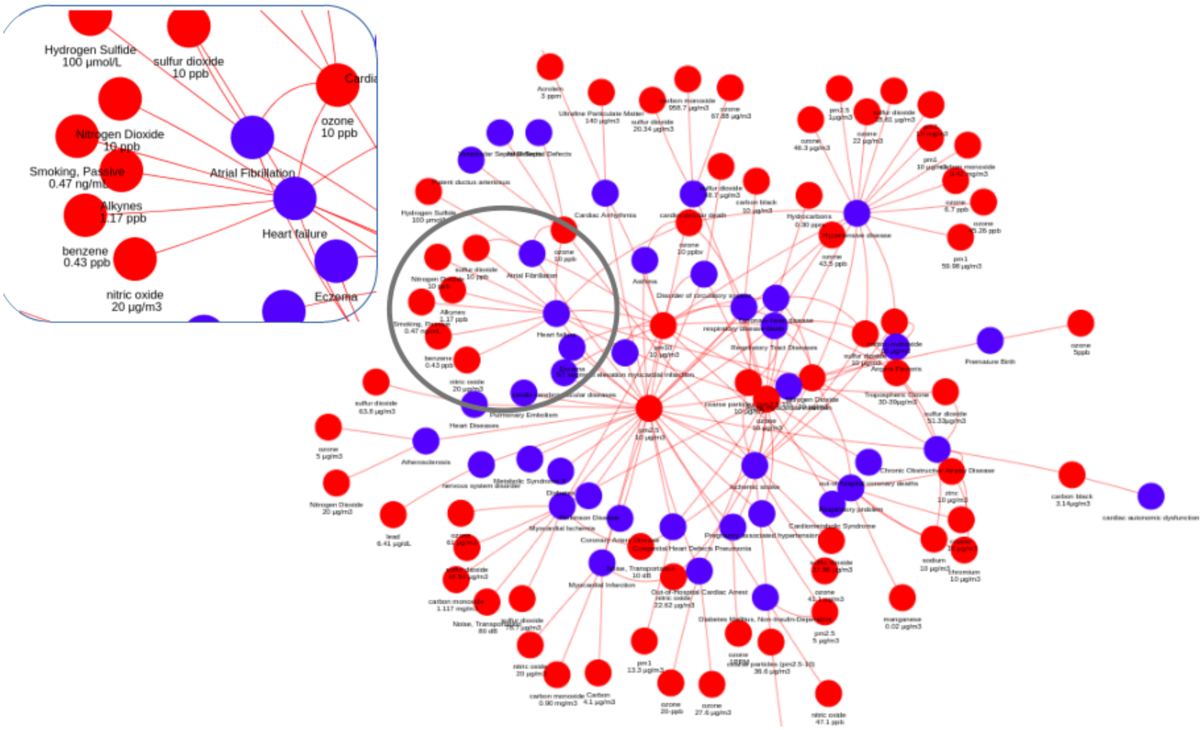


Figure 4: Snippet of 'PollCardioKG': pollutants are shown in red, diseases in blue, with edges representing causal relationships.

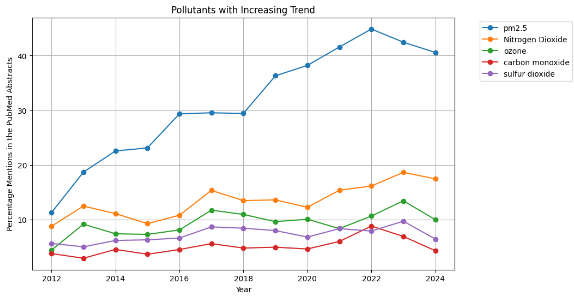
showing a clear upward trajectory over time. Hypertensive disease shows a significant rise, peaking around 2018, with consistently high mention rates thereafter. Ischemic Stroke and Atrial Fibrillation display gradual increases with slight fluctuations, while Cardiometabolic Syndrome and Diabetes maintain a steady but less pronounced rise, indicating continued relevance in medical research and public health.

Moreover, Figure 6(b) illustrates the trends of three specific diseases—Dyslipidemias, Metabolic Diseases, and ST-segment elevation myocardial infarction (STEMI)—in PubMed abstracts from 2012 to 2024. STEMI shows significant peaks in 2017 and 2020, particularly the highest occurrence in 2017, likely due to heightened research activity and updated clinical guidelines. Metabolic Diseases demonstrate notable increases in 2017 and 2020, followed by a consistent rise from 2021 to 2024, indicating growing recognition of their health impacts. In contrast, Dyslipidemias exhibit an irregular pattern with spikes in 2013, 2017, 2019, and a smaller rise in 2020, reflecting changes in research focus. These trends underscore the evolving landscape of disease research and the need for ongoing investigation into these conditions.

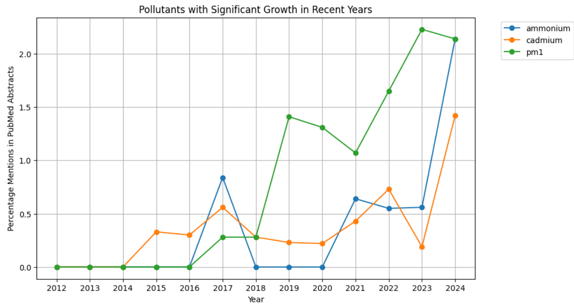
4.3 Trends of Pollutant Disease Correlation

Figure 7 highlights the increasing correlation between PM, the most frequently mentioned pollutants—and various diseases from 2012 to 2024 in PubMed abstracts. Both pollutants demonstrate a significant rise in association with health conditions such as hypertensive disease, cerebrovascular accidents, myocardial ischemia, and atherosclerosis, particularly after 2018 for $PM_{2.5}$. This trend emphasizes the growing awareness of the health risks linked to fine particulate matter, underscoring the necessity for continued research and public health initiatives to address the adverse effects of these pollutants on cardiovascular and cerebrovascular health.

Figure 8 presents a comprehensive visualization of the percentage of prevalent pollutant-disease correlations. In this heatmap, darker colors indicate stronger correlations, while lighter colors reflect weaker ones. Environmental Tobacco Smoke and Solid Fuel are notably associated with multiple diseases, including Myocardial Infarction and various Cardiovascular Events. Airborne Particulate Matter and Nitrogen Dioxide are significantly linked to conditions such as Cerebrovascular Accidents and Coronary Heart Disease. Additionally, pollutants like Cadmium and Lead show selective associa-

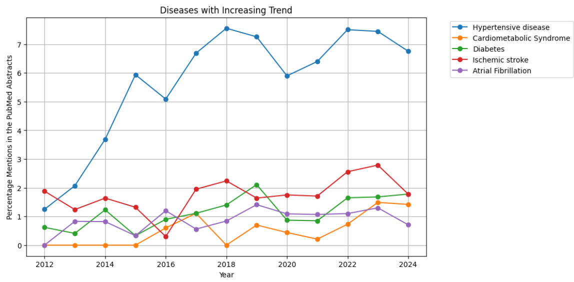


(a)

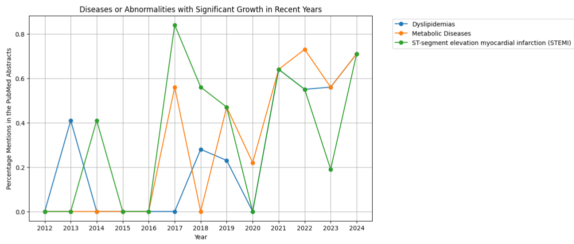


(b)

Figure 5: (a) Trends of the top five pollutants from 2012 to 2024. (b) Emerging pollutants in recent years.



(a)



(b)

Figure 6: (a) Trends of the top five diseases from 2012 to 2024. (b) Emerging diseases in recent years.

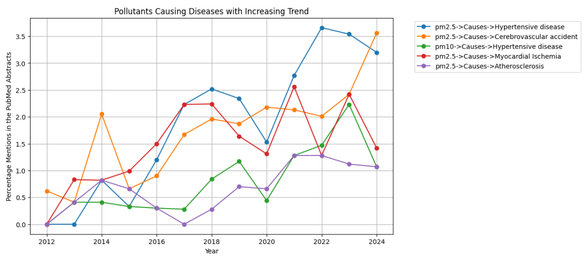


Figure 7: Most common diseases associated with PM exposure.

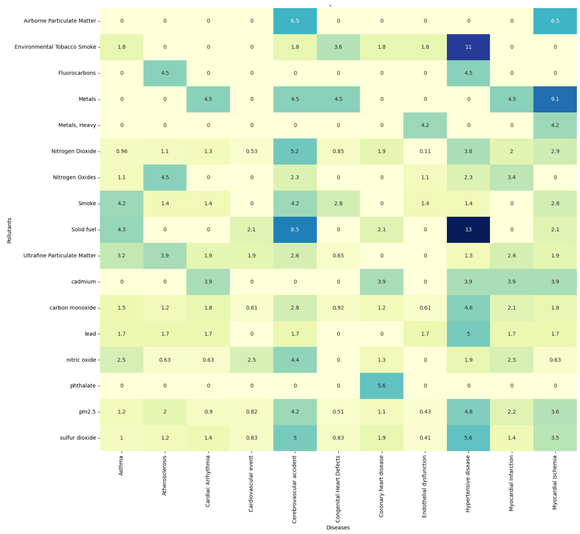


Figure 8: Heatmap illustrating the correlation between major pollutants and associated diseases.

tions with Endothelial Dysfunction and Myocardial Ischemia. These targeted correlations highlight the substantial health impacts of even less common pollutants, emphasizing the need for further investigation into their effects.

Furthermore, we explored the impact of varying concentration levels of the same pollutant on different cardiovascular diseases. Figure 9 illustrates the top five most commonly associated cardiovascular diseases (CVDs) linked to different levels of $PM_{2.5}$ exposure. Overall, this comprehensive analysis not only highlights critical pollutants that significantly affect various health conditions but also emphasizes the importance of addressing these environmental factors to enhance public health outcomes. By understanding these correlations, poli-

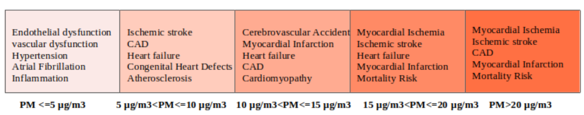


Figure 9: Top five most common CVDs associated with varying exposures to $PM_{2.5}$ concentrations..

cymakers and health professionals can better strategize interventions aimed at reducing exposure and mitigating associated health risks.

5 Conclusions

This study explores the causal relationships between environmental pollutants and cardiovascular diseases by constructing a comprehensive knowledge graph. Analyzing around 4,700 PubMed abstracts, we identified key pollutants and their associated health risks, including hypertensive disease and ischemic stroke. The knowledge graph reveals both direct and complex interrelations, offering a valuable tool for understanding the cardiovascular impacts of air pollution. Future work will integrate multimodal data to further enhance the analysis of indirect effects and expand the graph's scope for public health insights.

6 Limitations

This analysis relies solely on PubMed abstracts, which may lack detailed information on methodologies, pollutant concentrations, and demographic factors, potentially limiting the depth of insights. The use of specific keywords, such as "cardiovascular disease" and "air pollution", might omit relevant studies with alternative terminology. Furthermore, while correlations between pollutants and diseases are identified, determining true causality is difficult due to the potential influence of confounding factors like socioeconomic conditions and lifestyle behaviors.

References

Shaherin Basith, Balachandran Manavalan, Tae Hwan Shin, Chan Bae Park, Wang-Soo Lee, Jaetaek Kim, and Gwang Lee. 2022. The impact of fine particulate matter 2.5 on the cardiovascular system: a review of the invisible killer. *Nanomaterials*, 12(15):2656.

Rob Beelen, Massimo Stafoggia, Ole Raaschou-Nielsen, Zorana Jovanovic Andersen, Wei W Xun, Klea Katsouyanni, Konstantina Dimakopoulou, Bert Brunekreef, Gudrun Weinmayr, Barbara Hoffmann, et al. 2014. Long-term exposure to air pollution and cardiovascular mortality: an analysis of 22 european cohorts. *Epidemiology*, 25(3):368–378.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jing Feng and Wei Yang. 2012. Effects of particulate air pollution on cardiovascular health: a population health risk assessment. *PloS one*, 7(3):e33385.

Dieter Fensel, Umutcan Şimşek, Kevin Angele, Elwin Huaman, Elias Kärle, Oleksandra Panasiuk, Ioan Toma, Jürgen Umbrich, Alexander Wahler, Dieter Fensel, et al. 2020. Introduction: what is a knowledge graph? *Knowledge graphs: Methodology, tools and selected use cases*, pages 1–10.

Yanhui Jia, Zhennan Lin, Zhi He, Chenyang Li, Youjing Zhang, Jingyu Wang, Fangchao Liu, Jianxin Li, Keyong Huang, Jie Cao, et al. 2023. Effect of air pollution on heart failure: systematic review and meta-analysis. *Environmental Health Perspectives*, 131(7):076001.

Agnes Maria Lederer, Per Morten Fredriksen, Benedicta Ngwenchi Nkeh-Chungag, Frans Everson, Hans Strijdom, Patrick De Boever, and Nandu Goswami. 2021. Cardiovascular effects of air pollution: current evidence from animal and human studies. *American Journal of Physiology-Heart and Circulatory Physiology*.

Woojoo Lee, Youn-Hee Lim, Eunhee Ha, Yoenjin Kim, and Won Kyung Lee. 2022. Forecasting of non-accidental, cardiovascular, and respiratory mortality with environmental exposures adopting machine learning approaches. *Environmental Science and Pollution Research*, 29(58):88318–88329.

Mark R Miller and David E Newby. 2020. Air pollution and cardiovascular disease: car sick. *Cardiovascular research*, 116(2):279–294.

World Health Organization et al. 2016. Ambient air pollution: A global assessment of exposure and burden of disease.

Sung Kyun Park, Zhangchen Zhao, and Bhramar Mukherjee. 2017. Construction of environmental risk score beyond standard linear models using machine learning methods: application to metal mixtures, oxidative stress and cardiovascular disease in nhanes. *Environmental Health*, 16:1–17.

Sanjay Rajagopalan, Sadeer G Al-Kindi, and Robert D Brook. 2018. Air pollution and cardiovascular disease: Jacc state-of-the-art review. *Journal of the American College of Cardiology*, 72(17):2054–2070.

Peri L Schuyler, William T Hole, Mark S Tuttle, and David D Sherertz. 1993. The umls metathesaurus: representing different views of biomedical concepts. *Bulletin of the Medical Library Association*, 81(2):217.

Qinghua Sun, Aixia Wang, Ximei Jin, Alex Natanzon, Damon Duquaine, Robert D Brook, Juan-Gilberto S Aguinaldo, Zahi A Fayad, Valentin Fuster, Morton Lippmann, et al. 2005. Long-term air pollution exposure and acceleration of atherosclerosis and vascular inflammation in an animal model. *Jama*, 294(23):3003–3010.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ethan Waisberg, Joshua Ong, Mouayad Masalkhi, Sharif Amit Kamran, Nasif Zaman, Prithul Sarker, Andrew G Lee, and Alireza Tavakkoli. 2023. Gpt-4: a new era of artificial intelligence in medicine. *Irish Journal of Medical Science (1971-)*, 192(6):3197–3200.

Wenxuan Wang, Yihang Su, Jingyuan Huan, Jie Liu, Wenting Chen, Yudi Zhang, Cheng-Yi Li, Kao-Jung Chang, Xiaohan Xin, Linlin Shen, et al. 2024. Asclepius: A spectrum evaluation benchmark for medical multi-modal large language models. *arXiv preprint arXiv:2402.11217*.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2024. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, page ocae045.

Li Yujian and Liu Bo. 2007. A normalized levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):1091–1095.

Dong Zhang, Hanhan Li, Xiao-San Luo, Weijie Huang, Yuting Pang, Jinshan Yang, Mingwei Tang, Tariq Mehmood, and Zhen Zhao. 2022. Toxicity assessment and heavy metal components of inhalable particulate matters (pm2. 5 & pm10) during a dust storm invading the city. *Process Safety and Environmental Protection*, 162:859–866.

Yingying Zhang, Xiaotong Ji, Tingting Ku, Guangke Li, and Nan Sang. 2016. Heavy metals bound to fine particulate matter from northern china induce season-dependent health risks: a study based on myocardial toxicity. *Environmental Pollution*, 216:380–390.