Do Procedures Drive Prescriptions? A Quantitative Analysis Using the EHRShot Dataset

Starman4xz*
https://github.com/Starman4xz

GravityGravity*
https://github.com/GravityGravity

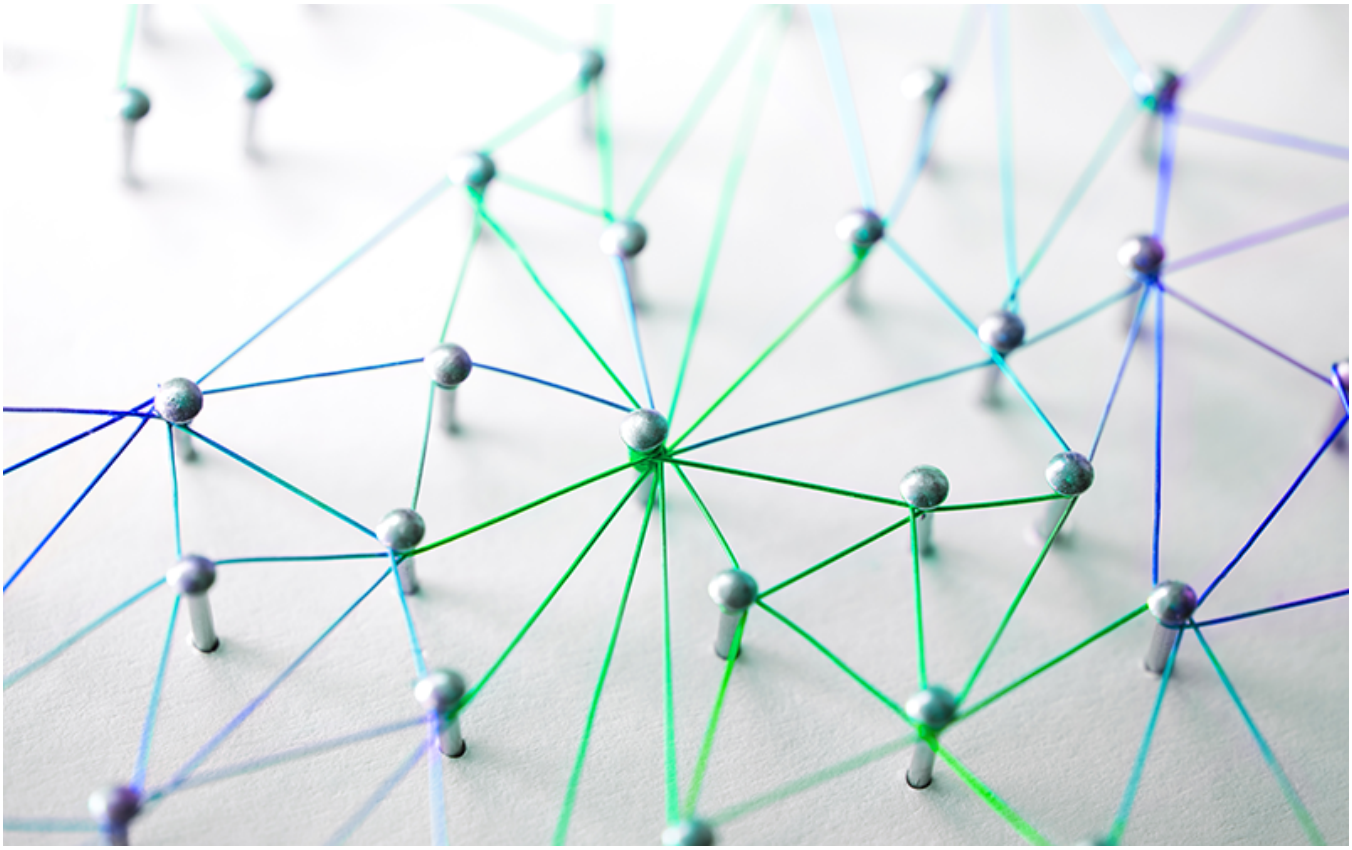

Figure 1: Sample representation of network cluster connections

Abstract

The increasing use of electronic health records (EHRs) provides new opportunities to study prescribing behaviors. In this study, we used the EHRShot data set to investigate whether medical procedures are associated with an increased probability of exposure to drugs. Using a reduced data set sample of the EHRShot records, we analyzed 2,000 patients using bipartite analysis methods. The results show that patients who underwent medical procedures were followed by a statistically significant increase in drug exposure over the course of their life. These findings suggest a strong link between procedures and increased in drug exposure.

1 Introduction

Network analysis has emerged as a powerful framework for examining complex relationships within healthcare data. In the context of electronic health records (EHRs), this approach enables the modeling of associations among diverse clinical entities such as diagnoses, procedures, and medications. Nodes in a healthcare network can represent patients or medical concepts, while edges capture interactions such as co-occurrence, temporal progression, or treatment dependencies. Network-based methods have been applied to identify comorbidity clusters among chronic diseases, characterize disease progression pathways, and analyze patient–drug or provider–patient relationships. By representing EHR data as interconnected systems, network analysis provides a scalable means to detect clinically relevant structures that are often hidden in traditional statistical models.

The present study addresses the question:

If a medical procedure was performed during the first six months of 2020, what was the likelihood of specific drug exposure?

The objective is to characterize the structural and statistical relationships between procedural interventions and subsequent medication use within this defined temporal window. The analysis seeks to determine whether certain procedures are consistently associated with increased prescription rates, to identify the drug classes most frequently linked to these procedures, and to assess the overall strength and direction of these associations across the examined population.

This research question held particular importance for understanding patterns of care and prescribing behavior within healthcare systems. Medical procedures often represent pivotal moments in patient treatment pathways. The downstream impact of procedures on subsequent medication exposure had not been sufficiently characterized. Quantifying the likelihood of drug exposure following procedural care provided insight into systematic prescribing tendencies, potential areas of overprescription, and opportunities for evidence-based guideline development. Such understanding is essential for improving patient expectations, optimizing resource allocation, and enhancing overall quality of care.

Previous research has leveraged EHR data to study medication patterns, drug exposure prediction, and healthcare utilization behaviors. Choi et al. (2020, Nature Scientific Reports) developed deep learning models to predict drug exposures using longitudinal patient data. Xu et al. (2019, Journal of Biomedical Informatics) analyzed co-occurrence networks to identify drug–disease associations, while Shang et al. (2021, BMC Medical Informatics and Decision Making) applied graph-based techniques to characterize comorbidity and treatment pathways. Although these studies have advanced understanding of drug utilization and disease interactions, most focus on limited therapeutic domains or specific disease categories. The broader, system-level relationship between medical procedures and subsequent prescription behaviors—particularly within a well-defined time frame—remains largely unexplored.

2 Network Construction Details

2.1 Dataset

The subset of data analyzed in this study was derived from the EHRShot dataset and restricted to patient visits that included both procedure occurrences and drug exposures during the same encounter within the first six months of 2020. Three primary tables were utilized: **Concept.csv**, sampled_procedure_occurrence, and sampled_drug_exposure.csv. The **Concept.csv** table provided the mapping between Athena vocabulary concept identifiers and their descriptive medical terms. The sampled_procedure_occurrence file contained information on procedure types, associated concept identifiers, and dates of performance, while the sampled_drug_exposure.csv file detailed drug exposures, drug types, and exposure start dates. Both the procedure and drug exposure files contained a **visit_id** column, which served as the key to align and match procedure events with drug exposures recorded during the same visit.

2.2 Preprocessing

All data preprocessing was performed in Python using built-in libraries for CSV handling to ensure reproducibility and transparency. The first filtering stage involved removing all procedure visits that had no corresponding drug exposures and all drug exposures that were not associated with any procedure visits. This was achieved by intersecting the **sampled_procedure_occurrence** and **sampled_drug_exposure.csv** tables on their shared **visit_id** column. The result produced two filtered files: one containing procedure visits with procedure identifiers, and another containing matched drug exposures or prescription identifiers. This filtering process reduced the dataset size by approximately a factor of ten, significantly improving computational efficiency.

A second filtering step was then applied to the **Concept.csv** file, which originally contained approximately 2 GB of vocabulary data. This file was filtered to retain only those concept identifiers corresponding to the procedure and drug IDs that were present in the filtered datasets. The resulting file was reduced to approximately 3 MB in size, containing only relevant mappings. To monitor the filtering process, the **tqdm** Python library was used to display progress bars and ensure successful execution of long-running operations, as local filtering scripts could take between five and ten minutes to complete depending on system performance.

Data standardization followed, including the removal of records with missing or null values in **visit_id** or **concept_id** columns, the conversion of all identifiers to consistent data types, and the verification of unique relationships between visits, drugs, and procedures. This step ensured that all remaining records were fully compatible for network construction.

2.3 Node Definition

Two types of nodes were defined in the constructed network:

(1) **Procedure nodes**, representing unique procedure_concept_id values from the `sampled_procedure_occurrence.csv` table.
(2) **Drug nodes**, representing unique `drug_concept_id` values from the `sampled_drug_exposure.csv` table.

Each node was assigned descriptive metadata based on its corresponding `Concept.csv` entry. Procedure nodes were assigned a *weight attribute* corresponding to the total number of occurrences of that procedure type across all visits in the first half of 2020. Drug nodes did not have associated weights, as the focus of weighting was on the frequency of procedural events.

2.4 Edge Definition

Edges were created between nodes when a procedure and a drug exposure shared the same `visit_id`. Each edge therefore represented a co-occurrence of a procedure and a drug within the same clinical encounter. It is important to note that this relationship captures *correlation, not causation*—the presence of an edge indicates that the drug was prescribed during a visit in which a given procedure was performed, but does not imply that the procedure directly caused the prescription.

Each edge was assigned a *weight* corresponding to the number of distinct visits in which that specific procedure–drug pair co-occurred. Thus, frequently co-occurring pairs received higher edge weights, allowing for identification of strong associations between particular procedures and medications.

For the purpose of the final network visualization, edges with weights less than 500 were removed to reduce visual clutter and focus on the most significant procedure–drug co-occurrences. This

thresholding did not affect the underlying network data used for quantitative analyses, but was applied solely for clarity in graphical representation.

2.5 Graph Construction and Implementation

Graph construction was performed using a custom Python script, `Graph_create.py`, developed for this project. The script utilized the `NetworkX` library to handle node and edge creation, attribute assignment, and weight computation. The filtered CSV files served as input to generate the bipartite network structure. NetworkX functions were used to add nodes for each unique procedure and drug identifier and to form edges based on shared `visit_id` values. Edge weights were accumulated during graph construction to reflect multiple co-occurrences across visits. The resulting graph object was stored in serialized form for downstream analysis, visualization, and computation of centrality and clustering metrics.

2.6 Graph Type

The resulting structure is best characterized as a **weighted bipartite network**, also referred to as a **two-mode heterogeneous graph**. One node set represents medical procedures, and the other represents drug exposures, with edges only existing between the two sets (i.e., no drug–drug or procedure–procedure edges). The inclusion of edge weights reflecting co-occurrence frequency classifies it further as a **weighted bipartite co-occurrence network**, suitable for subsequent projection or community detection analyses to explore relationships between procedures and drug utilization patterns.

3 Metrics and Analysis Methods

Network analysis was performed on two graphs derived from the EHRShot dataset: a **complete network** and a **filtered network**. The complete network included all recorded procedure–drug relationships, while the filtered network retained only high-confidence connections with substantial co-occurrence weights. Procedure nodes were labeled with the prefix "P-" and drug nodes with "D-" to distinguish entity types. Entries lacking valid concept identifiers were assigned the default values **P-0** or **D-0**, representing unmapped records. These zero-labeled nodes were included in the complete network for completeness but excluded from the filtered network to ensure that analyses reflected clinically meaningful associations. The nodes **P-0** and **D-0** were used as indicators of error, representing cases where procedures or drugs could not be identified; therefore, their corresponding edges hold no statistical meaning within the network analysis.

3.1 Selected Network Metrics

Several network metrics were applied to characterize structural properties and quantify associations between procedures and drug exposures. The selected metrics included **degree and weighted degree**, **network density**, **edge weight distribution**, **closeness centrality**, and **network comparison metrics**. Each metric provided a distinct perspective on the structure and connectivity of the networks.

Degree and Weighted Degree. Degree centrality measured the number of connections associated with each node, indicating how many distinct relationships a procedure or drug maintained within the network. Weighted degree (or node strength) extended this concept by summing the edge weights of each node, reflecting the cumulative frequency of co-occurrence. In the context of this study, procedures with high weighted degree values represented those most frequently associated with drug exposures, highlighting procedural types that were central to prescribing activity.

Network Density. Network density quantified the proportion of existing edges relative to all possible edges in the bipartite network. This metric captured the overall connectivity between procedures and drugs, serving as an indicator of how frequently procedures were linked to medication use. Comparing density between the complete and filtered networks provided insight into how the removal of unmapped or low-frequency nodes affected overall structural cohesion.

Edge Weight Distribution. The distribution of edge weights described how frequently procedures and drugs co-occurred within patient visits. Analyzing this distribution helped identify whether prescribing behavior was evenly distributed or dominated by a small subset of procedures and drugs. A heavy-tailed distribution suggested that a limited number of procedure–drug combinations accounted for most prescribing activity, while the majority occurred infrequently.

Closeness Centrality. Closeness centrality measured how close a node was to all other nodes in the network, based on the shortest path distances. Nodes with higher closeness values were more centrally positioned and could reach other entities with fewer intermediary connections. In this study, procedures or drugs with high closeness centrality represented those that were broadly integrated within the treatment network, indicating potential central roles in common care pathways.

Network Comparison Metrics. To assess differences between the complete and filtered networks, several comparative measures were calculated, including average degree, network density, and the size of the largest connected component. These metrics evaluated how data cleaning and filtering impacted network structure, providing a basis for interpreting the reliability and clinical significance of each representation.

4 Visualizations and Tables

Three network visualizations were generated to illustrate the structure and connectivity of the procedure–drug relationships derived from the EHRShot dataset. Figure 2 shows the **unfiltered network**, which includes all nodes and edges from the dataset. This graph captures the full scope of observed relationships but appears highly dense due to the inclusion of all recorded connections, including low-frequency and unmapped identifiers.

Figure 3 presents the **filtered network**, in which only edges with weights greater than 500 were retained. Node sizes for procedures were scaled according to their total frequency of occurrence to emphasize highly connected or commonly performed procedures. This visualization provides a clearer representation of dominant

procedural–pharmaceutical relationships while reducing noise from rare co-occurrences.

Finally, Figure 4 displays the **filtered network without the P-0 node**. The P-0 identifier corresponds to procedures lacking valid concept mapping and was excluded to improve interpretability. This version highlights clinically meaningful relationships by removing unmapped or incomplete records that do not contribute statistically to network structure.

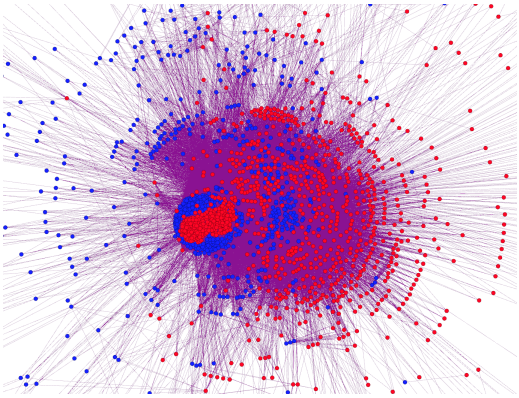

Figure 2: Unfiltered procedure–drug network containing all nodes and edges. ***Blue nodes represent procedures (P-) and red nodes represent drug exposures (D-).***

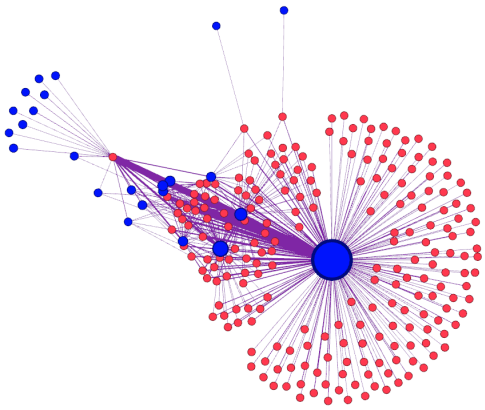

Figure 3: Filtered procedure–drug network (edges with weight > 500). Procedure node sizes reflect frequency of occurrence. ***Blue nodes represent procedures (P-) and red nodes represent drug exposures (D-).***

The scatter plot in Figure 5 illustrates the distribution of edge weights within the filtered procedure–drug network. Each point represents a unique procedure–drug pair, and its corresponding y-value indicates the frequency of co-occurrence within the same patient visit. The distribution shows a distinct right-skewed pattern, with the majority of edges concentrated below 1500 occurrences

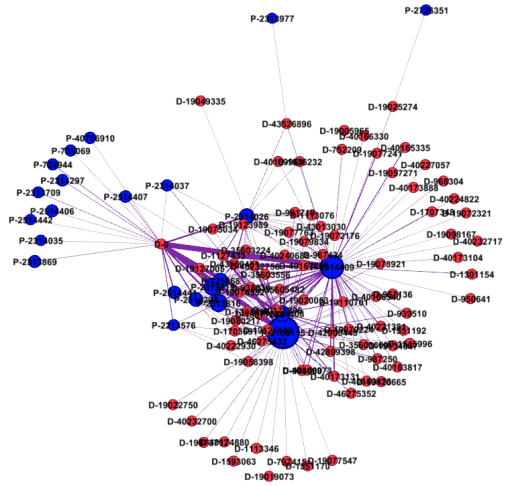

Figure 4: Filtered network excluding the P-0 node representing unmapped procedures. ***Blue nodes represent procedures (P-) and red nodes represent drug exposures (D-).***

and a small subset exceeding 2000. These high-weight edges represent the most frequent procedural contexts in which specific drug exposures occurred, highlighting dominant clinical interactions within the dataset. This pattern aligns with expected healthcare behaviors, where a limited number of procedures account for the majority of medication associations, while most procedure–drug relationships occur infrequently.

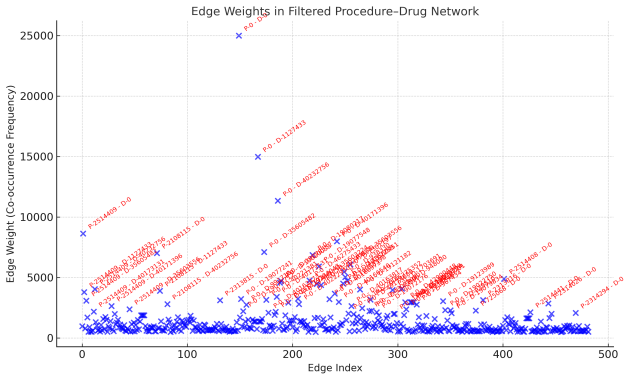

Figure 5: Scatter plot of edge weights for the filtered network, where each point represents a unique procedure–drug pair. Labels mark high-weight edges (weight > 2000) corresponding to the most frequent procedure–drug associations.

. The scatter plot of edge weights (Figure 5) served as the primary analytical tool for identifying the strongest procedure–drug relationships within the filtered network. By plotting the frequency of each procedure–drug co-occurrence, it was possible to visually distinguish a small subset of high-weight edges that stood apart from the dense cluster of lower-frequency connections. These outlier points correspond to procedure–drug pairs that occurred most

frequently across patient visits, forming the basis for the top ten relationships summarized in Table 1. Structurally, this pattern indicates a highly skewed network, where a limited number of procedural contexts dominate the overall landscape of medication exposure.

From a healthcare perspective, the prominence of a few high-weight edges reveals that medication use is not evenly distributed across all procedures. Instead, prescribing patterns are strongly shaped by recurring procedural activities—such as hospital care evaluations, blood draws, and electrocardiographic testing—that form the backbone of common clinical workflows.

Table 1: Top 10 most common procedure–drug pairs (fully mapped, excluding P-0 and D-0).

Proc. ID	Procedure Name	Drug ID	Drug Name	Weight
P-2514409	Subsequent hospital care...	D-40232756	oxycodone hydrochloride 5 MG Oral Tablet	4037
P-2108115	Collection of venous blood...	D-1127433	acetaminophen 325 MG Oral Tablet	3894
P-2514409	Subsequent hospital care...	D-1127433	acetaminophen 325 MG Oral Tablet	3793
P-2313816	Electrocardiogram, routine ECG...	D-35605482	2 ML ondansetron 2 MG/ML Injection	3077
P-2514408	Subsequent hospital care...	D-35605482	2 ML ondansetron 2 MG/ML Injection	2864
P-2314026	Pressurized or nonpressurized...	D-1127433	acetaminophen 325 MG Oral Tablet	2681
P-2313815	Electrocardiogram, routine ECG...	D-1127433	acetaminophen 325 MG Oral Tablet	2572
P-2108115	Collection of venous blood...	D-35605482	2 ML ondansetron 2 MG/ML Injection	2449
P-2514409	Subsequent hospital care...	D-35605482	2 ML ondansetron 2 MG/ML Injection	2206
P-2314026	Pressurized or nonpressurized...	D-40232756	oxycodone hydrochloride 5 MG Oral Tablet	2135

. Figure 6 illustrates that procedural activity within the network is highly centralized, with a limited number of procedures dominating the observed interactions. Each procedure node's frequency, represented by its node size in the overall network visualization, reflects how often that procedure co-occurred with at least one drug

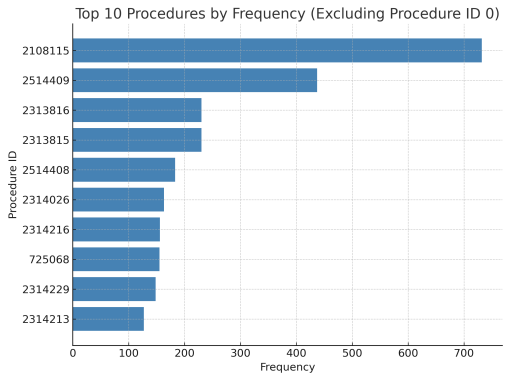

Figure 6: Top 10 most frequent procedure nodes after excluding unmapped identifier P-0. Each bar represents the number of times a specific procedure co-occurred with at least one drug exposure during the first half of 2020

exposure. This concentration reveals a hub-like network structure where a small set of procedural nodes drive much of the overall prescribing activity. In healthcare terms, this suggests that these common interventions underpin the majority of patient treatment events.

Table 2: Comparison of network metrics for Filtered vs. Complete graphs (first half of 2020).

Metric	Filtered	Complete
Nodes	97	2,065
Edges	244	106,538
Density	0.0524	0.0500
Average Degree	5.03	103.18
Average Weighted Degree	4,728.14	1,642.20
Average Closeness	0.4418	0.3904
Edge Weight (Min / Median / Mean / Max)	503 / 735.5 / 939.82 / 4,037	1 / 2.0 / 15.92 / 25,000

. Table 2 shows that the filtered graph—restricted to high-confidence, high-frequency procedure–drug links—has far fewer nodes and edges yet exhibits higher average weighted degree and higher average closeness, consistent with a more tightly connected core of clinically meaningful relationships. In contrast, the complete graph is comprehensive but dominated by low-weight edges (median edge weight 2.0), reflecting a long-tailed structure where many rare associations dilute centrality and connectivity. Interpreted in a healthcare context, the filtered network highlights routine procedural contexts that drive most drug exposure, whereas the complete network captures the broader, infrequent tail of prescribing activity that may be less impactful at the system level.

5 Interpretations and Conclusions

The structural patterns observed in the bipartite networks derived above reveal a distinct and meaningful relationship between medical procedures and subsequent drug exposures. The high connectivity and clustering of nodes suggests that procedures tend to be consistently followed by an increase in prescription events. Collectively, the network structure highlights that procedures are not

isolated events but rather integral points that trigger an increased likelihood of drug exposure for patients.

The calculated network metrics further reinforce the structural patterns observed in the visual analyses. The filtered graph, which isolates high-frequency and clinically relevant procedure–drug links, exhibits higher average weighted degree and greater closeness centrality compared to the complete network. This indicates that these connections form a tightly knit core of routine medical interactions where procedures and prescriptions are strongly interlinked. Conversely, the complete network has substantially more nodes and edges but a lower median edge weight, reflecting the presence of many weak, infrequent associations that dilute overall connectivity. The moderate network density across both graphs suggests a well-defined yet sparse structure typical of healthcare data, where specific clinical actions correspond to distinct medication profiles rather than universal prescribing patterns. Collectively, these metrics highlight that procedural care in healthcare systems is dominated by a concentrated set of strong, recurring relationships, emphasizing the central role of procedural workflows in shaping medication exposure patterns.

While informative, several limitations constrain the scope of the research. One such issue that could hinder the study is the data sparsity and sampling bias, as this research utilized a reduced sample size from the full EHRShot database, which may not fully capture all of the nuanced relationships between procedures and drug exposures. Additionally, the research data does not distinguish between preventative, therapeutic, or incidental drug prescriptions. Without the additional context, the motivations behind prescribing remain inferred rather than explicit.

In regards to future work, the research can build upon the foundational network analysis by incorporating **temporal network modeling** to capture whether procedures precede prescriptions, allowing for causality assessments to be made. Additionally, the analysis can be enhanced with the inclusion of **Graph Neural Networks (GNNs)** to analyze and predict future drug exposures based on patient or procedural history. These additions to the analysis will provide a better understanding of the connections between our research, and allow for improved validity of the report.

References

Received 13 September 2025; revised 26 Oct 2025; accepted 26 Oct 2025