Heliyon 10 (2024) e22766

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Quantitative patient graph analysis for transient ischemic attack risk factor distribution based on electronic medical records

Jian Wen [a,**,1], Tianmei Zhang [a,1], Shangrong Ye [a], Peng Zhang [a], Ruobing Han [a], Xiaowang Chen [a], Ran Huang [b], Anjun Chen [c], Qinghua Li [a,*]

[a] *Guilin Medical University Affiliated Hospital, Guilin, Guangxi, China*
[b] *West China Hospital, Chengdu, Sichuan, China*
[c] *Learning Health Community, Palo Alto, CA, USA*

ARTICLE INFO

Keywords:
Transient ischemic attack
Risk factor
Electronic medical records
Patient graph
Connection delta ratio
UMLS
Knowledge graph

ABSTRACT

A transient ischemic attack (TIA) affects millions of people worldwide. Although TIA risk factors have been identified individually, a systemic quantitative analysis of all health factors relevant to TIA using electronic medical records (EMR) remains lacking. This study employed a data-driven approach, leveraging hospital EMR data to create a TIA patient health factor graph. This graph consisted of 737 TIA and 737 control patient nodes, 740 health factor nodes, and over 33,000 relations between patients and factors. For all health factors in the graph, the connection delta ratios (CDRs) were determined and ranked, generating a quantitative distribution of TIA health factors. A literature review confirmed 56 risk factors in the distribution and unveiled a potential new risk factor "rhinosinusitis" for future validation. Moreover, the patient graph was visualized together with the TIA knowledge graph in the Unified Medical Language System. This integration enables clinicians to access and visualize patient data and international standard knowledge within a unified graph. In conclusion, graph CDR analysis can effectively quantify the distribution of TIA risk factors. The resulting TIA risk factor distribution might be instrumental in developing new risk prediction machine learning models for screening and early detection of TIA.

1. Introduction

Transient ischemic attack (TIA) increases the risk of stroke, which affects 15 million people worldwide each year [1,2]. Recent analysis of the long-term population-based Framingham Heart Study in the US from year 1948–2017 has established that the TIA incidence rate is approximately 1.19/1000 person-years and, as a stroke risk factor, TIA has an adjusted hazard ratio of 4.37 [3]. After an initial TIA incidence, the risk of recurring TIA, stroke or death varied between 6 and 30 % depending on the populations and the time period [3–5].

Recognizing and treating TIA can lower the risk of a major stroke [6]. However, in developing countries, public awareness of TIA is very limited. For example, only about 3 % of adults have knowledge of TIA in China [7]. As a result, TIA is predominately undiagnosed and untreated within the country. There is a pressing need to enhance the detection and suitable management of TIA.

* Corresponding author. Dept. of Neurology Guilin Medical University Affiliated Hospital, 15 Lequn Road Guilin, Guangxi, 541000, China.
** Corresponding author. Dept. of Neurology, Guilin Medical University Affiliated Hospital, 15 Lequn Road, Guilin, Guangxi, 541000, China.
E-mail addresses: wenjian2400@163.com (J. Wen), qhli1999@163.com (Q. Li).
[1] JW and TZ contributed equally to this work and both were considered the first author.

https://doi.org/10.1016/j.heliyon.2023.e22766
Received 10 October 2022; Received in revised form 26 October 2023; Accepted 19 November 2023
Available online 25 November 2023

The understanding of TIA risk factors is foundational for TIA prevention, early detection, diagnosis, treatment, and management [8]. TIA shares most of the risk factors with stroke. In stroke and TIA patients, no major difference in risk factors was found when comparing individuals younger than 50 years to older patients [9,10]. Several risk factors are still under investigation. Studies and case reports from the US [11,12], UK [13], Korea [14], and China [15,16] have suggested rhinosinusitis or sinusitis as potential risk factor for stroke. However, no study has yet reported an association between rhinosinusitis and an increased risk of TIA.

In this study, a risk factor is defined as a health factor that negatively affects one's health, such as causing or contributing to a disease. There are different methods to study disease risk factors. The traditional method to determine a risk factor for a disease is a hypothesis-driven association study, i.e., designing a prospective cohort study or retrospective case-control study for a given factor in question and calculating the risk ratio or odds ratio from the study data by single variate and/or multivariate statistical analysis [8]. The genomics era which arrived about 20 years ago has given rise to a data-driven approach for genome-wide association studies aimed to study multiple genetic variants as influencing factors of target disease(s) simultaneously [17,18]. In recent 10 years, as electronic medical record (EMR) system has become ubiquitous in hospitals, machine learning (ML) using EMR-wide variables also represents a data-driven approach for studying risk prediction models [19,20].

Inspired by these advancements in methodology, we have been interested in exploring a data-driven approach to study risk factors for different diseases, such as TIA, stroke and lung cancer, using all of the available health factors collected from EMR patient data. One recently published graph method was able to create risk factor distribution and reveal potential new risk factors of disease from synthetic patient data [21]. This data-driven method is complementary to the traditional hypothesis-driven method designed for determining the associations of specific disease risk factors.

The Unified Medical Language System (UMLS), an international standard biomedical knowledge graph (KG), integrates all of the main international standard terminologies, classifications and coding standards [22]. With the recent graph technologies, one can conveniently construct and analyze KGs [23,24]. For example, integration of UMLS KG and patient graphs enabled patient data semantic search as well as clinical decision support [25]. A MIT study showed that patient symptom knowledge graphs made from basic concept extractions of EMR data were able to predict clinical diagnosis [26].

Although patient data graph representation is promising in generating new insights gleaned from EMR data and transforming such insights into health care actions, it is still limited using EMR-wide graphs to study disease risk factors, diagnoses, and treatments [27].

Using a hospital EMR data-driven approach, this study aimed to construct TIA patient graphs and create TIA risk factor distribution, as well as integrate the quantified risk factors with the international standard UMLS knowledge graph. The results may have significant implications for TIA risk prediction, early detection and management that are based on better understanding of TIA risk factors.

2. Materials and methods

2.1. Data collection from EMR

This study utilized patient data from EMR and received approval from the IRB of Guilin Medical University Affiliated Hospital (QTLL202139) in China. EMR records spanning January 2018 to June 2021 were de-identified and saved on a secure data server overseen by the hospital's informatics department. The dataset comprised approximately 1 million patients and 7 million encounters, both outpatients and inpatients. Personal details such as patient names, birth dates, addresses, and contact information were eliminated. The original numbers for patients and encounters were substituted with random, unrelated numbers. Prior to data usage, our research team underwent training on the hospital's data security and patient privacy policy.

Due to the absence of standard diagnostic codes within the EMR data, synonyms for TIA in Chinese were employed to identify TIA patients. The target dataset included 737 TIA patients aged 30 and above. Concurrently, 1448 patients of similar age but without a TIA diagnosis were chosen as controls.

De-identified records of patient visits, diagnoses, laboratory tests, and procedures were incorporated into a data collection tool on the secure server. The EMR data was classified by researchers into nine categories: history, health factor, risk factor, medication, lab test, observation, condition, symptom, and treatment. We used a semi-automated data collection method. The lab test data, owing to its uniform format, was auto-extracted by our tool and stored in the database. However, the data from the remaining eight categories, due to their complexity, required manual input by the researchers. For instance, a "History of hypertension" was classified in the "history" category, while "Physical weakness on one side" was categorized under "symptom." This semi-automatic method optimized the efficiency of our data collection. Given the uncoded nature of the records, pragmatic rules were devised to bolster consistency in collecting data. Synonyms underwent conversion to "local standard terms," resulting in what we termed "local standard data." For instance, phrases like "walk slowly" and "walk unsteadily" were both translated as "walk with difficulty," while "Left upper limb weakness" and "Weakness in the lower right limb" were collectively rendered as "Physical weakness on one side".

Subsequently, we established a standardized database with categorized data. Only the data proceeding the final TIA diagnosis of each patient were considered for the study of disease risk factors. A Patient Diagnosis Journey (PDJ) object encapsulated one or multiple encounters leading to the final diagnosis. Upon exporting PDJ data into a CSV file for analysis, only the most recent data for each health factor in the PDJ was chosen. The concluding raw dataset contained nearly 14,000 data items from TIA patients and around 50,000 data items from control patients. Over 3000 distinct health factors were pinpointed within these datasets. For instance, a patient's data item might encompass the aforementioned nine categories, where the "History of hypertension" is under the "history" category and a symptom like unilateral limb weakness is categorized under "symptom."

2.2. Building patient health factor graphs

To streamline the patient graph, we converted continuous numeric data into categorical data. For instance, age values were categorized into ranges: 30–50, 50–70, and >70 years old. Drinking was labeled as "true" if consumption exceeded 2 drinks per day, and smoking was considered "true" for those consuming one or more cigarettes daily. Lab test results from the EMR were already categorized as: normal/abnormal, true/false, positive/negative, high/medium/low, up/down/normal. Following this data conversion, an equal number (737) of background patients were randomly selected. Both the TIA and background factor data, amounting to approximately 33,000 standard data points across roughly 630 factors (i.e. codes), were consolidated into a factor import file of CSV format: virtual-id, category, code, term, value, unit, converted-value, date. Both TIA and background patients (a total of 1474) were cataloged in a patient import file, with each line representing a patient with format: virtual-id, TIA-label (1 for TIA, 0 for background), factor-count.

We employed the Neo4j Desktop tool (version 4.4), freely available from Neo4j Inc. (San Mateo, California, USA), for constructing patient graphs. Neo4j features a graphical user interface (Neo4j Browser) to facilitate queries using the Cypher language and to visualize graphs. The database offers an application programming interface (API) for Python driver to import CSV files to formulate graphs. Within our patient-centric graph model, each patient was denoted by a "Patient" node (1474 in total). Pairs of health factors and their corresponding values were represented by approximately 740 factor nodes. Given that all values were categorized and some factors have multiple categorical data points, the count of factor-value pair nodes grew from about 630 to around 740. Depending on a factor's category, the factor node was labeled as one of the following: Condition, Symptom, Observation, History, RiskFactor, Labtest, Procedure, Medication, or Treatment. The graph charted over 33,000 connections from patients to various factors. Constraints were set for each label to guarantee their uniqueness. Patient nodes needed a virtual-id while all factor nodes demanded a category, code, and converted-value as node keys.

2.3. Generating distribution of risk factors by graph analysis

We utilized the graph method detailed in the synthetic data study [25]. A python script was employed to automatically query the patient graph for each health factor. We separately counted the number of connections from each factor to both the TIA target patients (TPC) and background patients (BPC) that appeared in the search results. For each factor, we calculated a connection delta ratio (CDR) as (TPC – BPC)/(TPC + BPC). This CDR serves as a relative measure of the strength of connections from a factor to the target patients. A CDR value between 1 and 0 indicated that the factor is more strongly associated with the target patients than with the background patients. The higher the value, the stronger the association. Conversely, a CDR below 0 indicates that the factor is more related to the background patients. By sorting the health factors based on their CDR values and plotting these values against the sorted factors, we established a distribution of TIA health factors ranked from highest to lowest strength.

In this study, we selected factors that had a CDR greater than a 0.1 cutoff and were connected to more than 10 TIA patients for literature verification. We conducted literature searches using English terms translated from the local standard terms, aiming to verify whether a health factor is a confirmed risk factor. If a factor's association with TIA was inconclusive in the literature, we labeled it as "unsure". If an "unsure" factor exhibited a high CDR and was connected to a significant number of TIA patients (more than 50), it was marked as "cdr-suggested." This label implies that the factor, as suggested by the CDR analysis, may be a potential new risk factor warranting further research.

2.4. Review of diagnostic images for TIA and rhinosinusitis

We selected 48 TIA patients from the same hospital for a review of their TIA-related imaging results from January to June 2022. The imaging modalities evaluated included head computed tomography (CT), head magnetic resonance imaging (MRI), neck computed tomography angiography (CTA), whole-brain CTA, ultrasonography of the head and neck, and magnetic resonance angiography (MRA). For TIA, presence of intracranial and extracranial carotid artery stenosis and plaque was assessed. A clinician evaluated the presence or absence of rhinosinusitis using the head CT, head MRI, neck CTA and MRA images.

2.5. UMLS TIA biomedical factor KG

The 2020AB release of UMLS data was downloaded from the UMLS website [26]. It was installed locally, and the common terminology sources were selected, resulting in a knowledge base containing about 2.8 million concepts with concept unique identifiers (CUI). These concepts were represented by about 8.3 million terms and had about 39.1 million relationships in rich release format (RFF). To construct a UMLS TIA KG, the target diseases were first expanded from TIA to all its child concepts in the UMLS hierarchy. The biomedical concepts horizontally related to the target disease concepts were retrieved from the concept relations file MRREL. RFF using a list of selected relationship attributes (RELA), which excluded the parent/child and other relationships less relevant to disease biomedical factors. These relationships were categorized into two groups: biological relationship ("biorel") and medical relationship ("merel"). The TargetConcept nodes were first connected to the RelCat nodes, which were then connected to the related Concept nodes. An AbstractPatient node was linked to the TargetConcept nodes to complete the TIA biomedical factor knowledge graph.

2.6. Integration of TIA patient graph and UMLS TIA KG

We employed two methods for graph integration. In the first method, an individual patient node from the EMR patient graph was directly linked to the AbstractPatient in the UMLS KG. In the second method, risk factors based on distribution were introduced as Concept nodes into the UMLS TIA KG. These risk factors were grouped under a third category labeled as a risk factor relationship ("rfrel") and were linked to a RelCat node. The headers for the risk factor import file included: cat, codevalue, code, factor, value, CDR, and tag.

3. Results

3.1. TIA patient health factor graph

To study risk factors of TIA in graph, we designed a patient-centric graph model with patient nodes linked to health factor nodes that represented all health factors in different categories (Fig. 1). The patient-factor relationship was labeled according to the category of the factor node (Table 1). After data were imported into the Neo4j database, the patient graph contained 737 TIA patients and 737 background patients with about 740 health factor nodes and over 33,000 relations between patients and factors.

The first patient graph analysis was graph search with Cypher queries listed in Table 2, presenting patient topologies in relation to any number and type of risk factors. An example topology of patients and 5 co-occurring diseases is shown in Fig. 2a. Fig. 2b shows an example topology of patients with 1 medical history and 5 symptoms, while Fig. 2c for patients with 1 lab test and 4 imaging observations. These topologies provide visualization of patient and factor relationships.

3.2. Distribution of TIA risk factors

For the second analysis of patient graph, we quantified TIA risk factors based on the topologies of search results for each factor. This allowed us to produce a distribution of risk factors and to uncover potential new risk factors that might have been previously overlooked. The patient graph was automatically searched for each factor and its respective value. CDRs were determined by comparing the number of connections to TIA patients against the number of connections to background patients. Since the CDR represents the relative strength of a factor's association with TIA – essentially indicating its relatedness - sorting the factors by CDR in descending order created a distribution of TIA health factors. To further analyze risk factors, we examined the top factors ranked by CDR above a designated CDR cutoff of 0.1. We excluded factors that didn't contribute to the risk of TIA and validated the remaining factors against existing literature.

Table S1 lists the resulting distribution of TIA risk factors, which was in very good agreement with the literature as most of the risk factors were confirmed [8]. The risk factor distribution curve is shown in Fig. 3. The top CDR-ranked risk factors included co-occurring

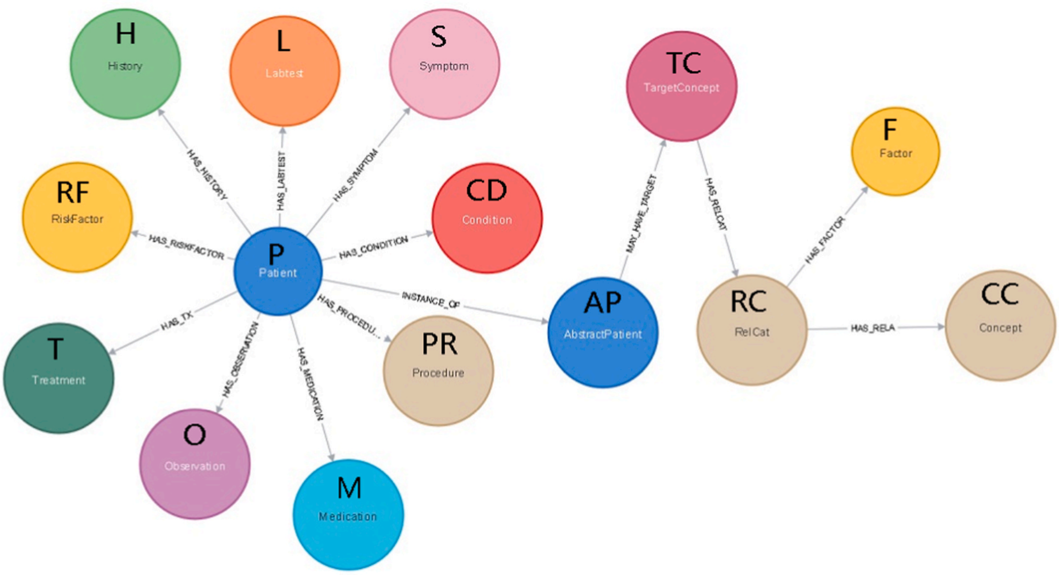

Fig. 1. Integrated graph model for health factors of TIA patients and biomedical concepts in UMLS TIA subgraph. On the left: Patient node (blue) is linked to various categories of health factors. On the right: AbstractPatient node (blue) is linked to UMLS TargetConcept node for disease, which is linked to RelCat (relationship category) node with either Concept node, Factor node, or both connected. Refer to Table 1 for the node labels and relationship labels.

Table 1
Graph node labels and relationship labels. Node letters from Fig. 1. Three relationship categories (RelCat): biorel, medrel, and rfrel.

From-Node: Label	Relationship Label	To-Node: Label
P: Patient	HAS_CONDITION	CD: Condition
P: Patient	HAS_SYMPTOM	S: Symptom
P: Patient	HAS_HISTORY	H: History
P: Patient	HAS_OBSERVATION	O: Observation
P: Patient	HAS_RISKFACTOR	RF: RiskFactor
P: Patient	HAS_PROCEDURE	PR: Procedure
P: Patient	HAS_MEDICATION	M: Medication
P: Patient	HAS_TREATMENT	T: Treatment
P: Patient	HAS_LABTEST	L: Labtest
P: Patient	INSTANCE_OF	AP: AbstractPatient
AP: AbstractPatient	MAY_HAVE_TARGET	TC: TargetConcept
TC: TargetConcept	HAS_RELCAT	RC: RelCat
RC: RelCat	HAS_RELA	CC: Concept
RC: RelCat	HAS_FACTOR	F: Factor

Table 2
Example graph search tasks and queries. Queries in Cypher language were used to search the integrated graph of patient health factors and UMLS concepts. Label '1' for lung cancer patient, 0 for control patient. C-number: local code.

No	Example Search Tasks	Cypher Search Queries
1	Search patients with 5 co-occurring diseases: <p class="">C-746982 Hyperhomocysteinemia C-690743 Epilepsy C-539246 Rhinosinusitis C-845276 Hyperlipidemia C-649035 Hypertension	match (p:Patient {label:'1'})→(f) where f.code = 'C-746982′ or f.code = 'C-690743′ or f.code = 'C-539246′ or f.code = 'C-845276′ or f.code = 'C-649035′ return p, f;
2	Search patients with 1 medical history and 5 symptoms: C-564918 History of taking blood pressure C-254917 Speaking impairment C-841063 Physical weakness on one side C-938176 Numbness on one side C-183659 Memory loss C-310857 Double vision	match (p:Patient {label:'1'})→(f) where f.code = 'C-564918′ or f.code = 'C-254917′ or f.code = 'C-841063′ or f.code = 'C-938176′ or f.code = 'C-183659′ or f.code = 'C-310857′ return p, f;
3	Search patients with 5 lab tests and observations: C-684521 Homocysteine C-435769 Carotid plaque C-536280 Cerebral artery stenosis C-391827 One-side vertebral artery stenosis C-713869 Sinus cyst	match (p:Patient {label:'1'})→(f) where (f.code = 'C-684521′ and f.valcvt = 'up') or (f.code = 'C-435769′ and f.valcvt = 'true') or (f.code = 'C-536280′ and f.valcvt = 'true') or (f.code = 'C-391827′ and f.valcvt = 'true') or (f.code = 'C-713869′ and f.valcvt = 'true') return p, f;

diseases, such as hyperhomocysteinemia, hyperlipidemia, hypertension and diabetes; biomarker homocysteine; medical histories of blood pression control, antiplatelet medication and arterial stenting; imaging observations of stenosis, plaque, arteriosclerosis, and abnormal electroencephalogram (EEG) [28]. The CDR analysis also found known symptomatic risk factors like weakness, numbness, dizziness, speech impairment, memory loss, double vision as well as lifestyle risk factors like smoking. The known TIA mimic – epilepsy was also identified [29].

Two unexpected factors were found by CDR in the distribution: rhinosinusitis and tinnitus. Tinnitus was reported as a risk factor in young adults in one case-control study in Taiwan, China [30] and some cases in the US [31], but whether it was considered a risk factor was inconclusive. Since there were only 13 TIA patients appeared to have tinnitus in our dataset, the status of tinnitus was tagged "unsure".

Several studies have associated rhinosinusitis with stroke [11–16], but no report has specifically linked it to TIA. Our CDR analysis suggested "rhinosinusitis" imaging observation to be a potential TIA risk factor (tagged "cdr-suggested"). Two other factors "Sinus cyst" and "Deviated nasal septum" in the distribution were observations related to rhinosinusitis and thus tagged "cdr-suggested". The observation of rhinosinusitis was verified by our clinicians reviewing the imaging results of new TIA patients in the first 6 months of 2022. Out of 48 TIA patients, 24 patients (50 %) had images with rhinosinusitis observed along with stenosis and/or plaque found in blood vessels. However, it requires further association studies to determine whether rhinosinusitis is a risk factor for TIA.

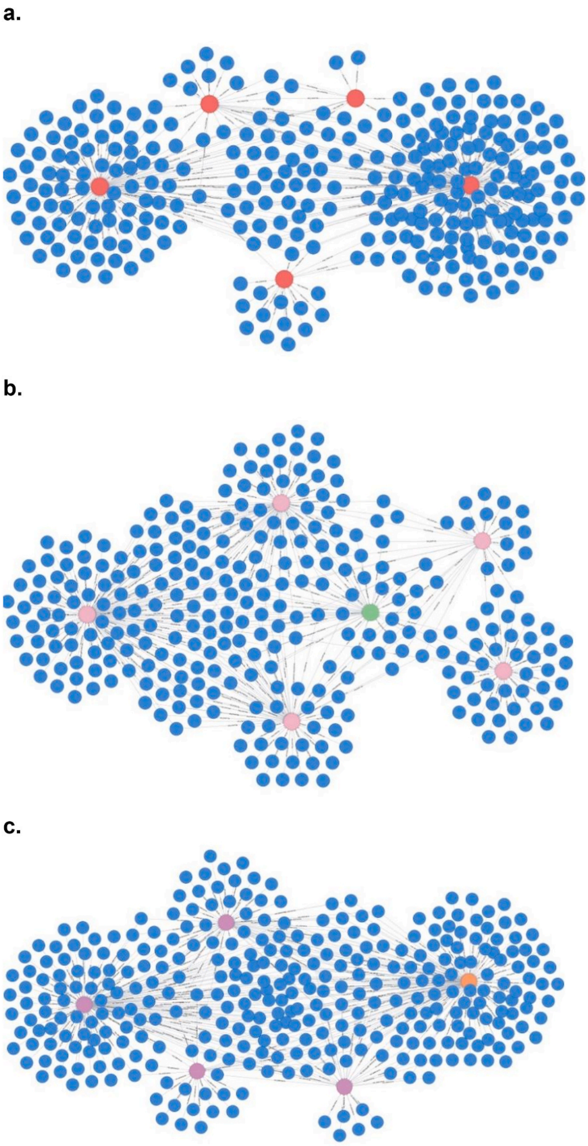

Fig. 2. Topologies of example TIA patient graphs resulted from graph searches with different numbers and categories of health factors. Cypher search queries in Table 2 **a).** Query #1 with 5 co-occurring diseases (red). **b).** Query #2 with 1 medical history (green) and 5 symptoms (pink). **c).** Query #3 with 1 lab test (yellow) and 4 observations (pink). Blue node: patient.

3.3. Integration of TIA patient graph with UMLS TIA biomedical KG

The TIA concept was expanded to a set of 23 concepts including immediate children of TIA in the UMLS disease hierarchy. Only in 4 of these concepts we found 14 biomedical relations with the desired relationship attributes (Table 3). As shown in Fig. 4, from one individual TIA patient's health factor graph, this small UMLS TIA biomedical knowledge graph with 11 related medical concepts and 3 related biological concepts was brought into an integrated view. This was the third patient graph analysis that was able to conveniently bring the international standard biomedical knowledge into the clinical workflow.

3.4. Integration of TIA risk factors into UMLS TIA KG

It was clear that the UMLS TIA KG lacked risk factor relations. We integrated the TIA risk factors identified above (Table S1) into the UMLS TIA KG to fill the gap. After the integration, the UMLS TIA biomedical knowledge graph became richer in risk factor contents as shown in Fig. 5. This result demonstrated that the distribution of TIA risk factors could be a new source of risk factor data for anyone to integrate into UMLS KG. The risk factor nodes in the integrated KG were notably different from the UMLS concept nodes. A quantified risk factor node represented a pair of health factor concept and its respective value, and it was quantified with CDR for relative

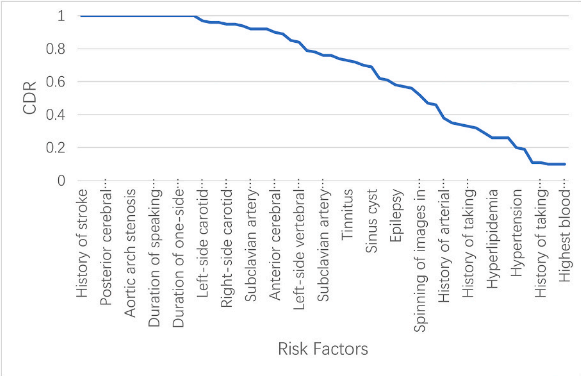

Fig. 3. Distribution curve of risk factors of TIA as sorted by the connection delta ratio (CDR). Only some factors are shown on the x-axis. Refer to Table S1 for the complete list of TIA risk factors.

Table 3
Horizontally related biomedical concepts for TIA in the UMLS TIA knowledge graph. CUI: concept unique identifier. CUI1: target disease CUI including the first level of child concepts that have horizontal relations. CUI2: related biomedical concept CUI. REL: UMLS relationship. RELA: UMLS relationship attributes.

CUI1 Term1	REL, RELA, RelCat	CUI2 Term2
C0007787 Transient Ischemic Attack	RO, may_prevent, medrel	C0004057 aspirin
	RO, may_prevent, medrel	C0043031 warfarin
	RO, may_prevent, medrel	C0282378 warfarin potassium
	RO, may_prevent, medrel	C0376218 warfarin sodium
	RO, may_prevent, medrel	C0981808 acetylsalicylate sodium
	RO, is_primary_anatomic_site_of_disease, biorel	C0006104 Brain
	RO, has_associated_finding, medrel	C0455536 History of transient ischemic attack
	RO, has_associated_finding, medrel	C0475701 Family history of transient ischemic attack
	RO, has_associated_finding, medrel	C3532623 Suspected transient ischemic attack
C0038531 Subclavian Steal Syndrome	RO, associated_morphology_of, biorel	C0028778 Obstruction
	RO, associated_morphology_of, biorel	C1261287 Stenosis
C1960656 Transient cerebral ischemia due to atrial fibrillation	RO, cause_of, medrel	C0004238 Atrial Fibrillation
C4039815 Transient ischemic attack due to embolism	RO, cause_of, medrel	C0007780 Cerebral Embolism
	RO, has_associated_finding, medrel	C4039272 History of transient ischemic attack due to embolism

strength.

4. Discussion

This study represented the first quantitative graph analysis for TIA risk factor distribution, and it has achieved the two study goals: (1) Created the distribution of TIA risk factors ranked by the relative strength measure CDR, and (2) Integrated the quantified TIA risk factors into the UMLS TIA biomedical knowledge graph, filling the gap in standard KG. The study results may have significant implications for development of TIA risk prediction machine learning models, TIA screening and early detection, and disease management that requires comprehensive understanding and quantification of risk factors [32].

The finding of rhinosinusitis as a potential new risk factor demonstrated that the data-driven graph CDR analysis method can reveal potential new risk factors in the risk factor distribution generated from EMR data. The assignment of "cdr-suggested" status of a health factor depends on the amount of data. Improving the reliability of the assignment requires collecting data from more target patients. It is worth emphasizing that conducting retrospective case-control studies or prospective clinical trials is required to determine whether rhinosinusitis is directly associated with TIA. Validation of rhinosinusitis as TIA risk factor would be a future research direction.

As demonstrated, patient graphs can be built from EMR data to enable graph search on patient data of an entire hospital. It can also

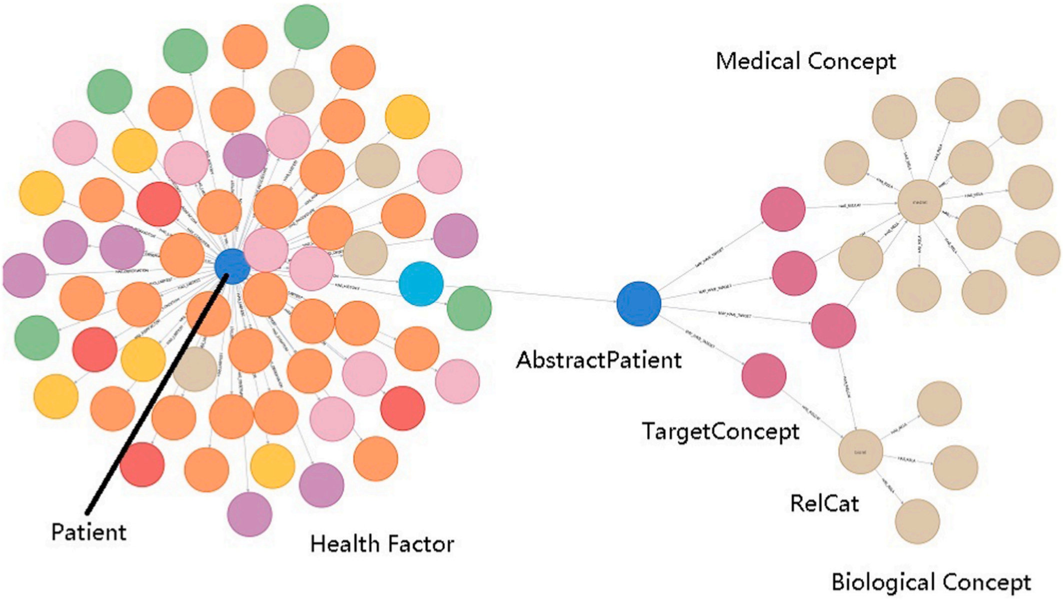

Fig. 4. Connecting one example TIA patient graph to UMLS TIA biomedical knowledge graph. TargetConcept nodes (pink): TIA and child diseases. RelCat (gray): biorel, medrel.

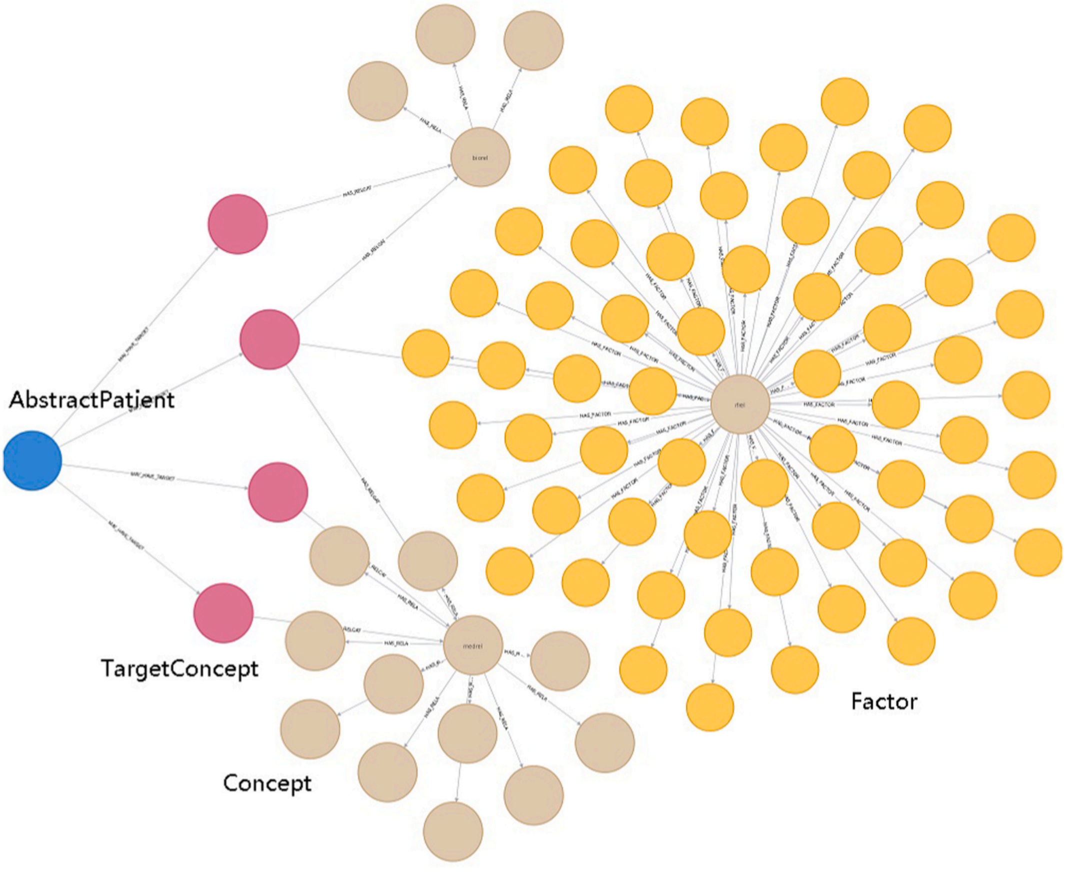

Fig. 5. Integration of TIA risk factors into UMLS TIA KG. TargetConcept nodes (pink): TIA and child diseases. RelCat (gray): biorel, medrel, rfrel. Factor nodes (yellow): quantified risk factors.

serve as a convenient way to bring relevant biomedical knowledge from the international standard UMLS KG to clinical workflow. This graph integration makes it possible for clinicians to search and visualize patient graphs from EMR and standard knowledge graphs from UMLS at the same time, which may present graph views of patient care complementary to the traditional table view.

The patient graph analysis method used in this study is generally applicable to other diseases and conditions. From a hospital's EMR, a quantitative distribution of health factors can be generated for each disease. With a thorough literature review, this distribution can establish a quantitative distribution of known risk factors for feature engineering in machine learning. This distribution can also reveal previously unknown potential risk factors for future verification studies.

One important use of the risk factor distribution is for machine learning of EMR data. Although there are many ML model studies for stroke risk prediction [33,34], studies on developing ML models to predict TIA risk were very limited. Bacchi et al. built a convolutional neural network model on clinical notes, which predicted TIA-like presentation with an AUC of 0.819 or 0.883 if data included MRI reports [35]. It would be another future research direction to apply TIA risk factor distribution in feature engineering of ML studies using EMR data.

Development using EMR-wide data has some limitations. Special attentions should be given to missing data and data bias in EMRs. Outpatient records usually have fewer data points compared to inpatient records. For example, the outpatient encounters had little data in medical history information, symptom records, laboratory data and so on. Different physicians may record patient data very differently, causing data variations for the same disease. For example, symptoms such as "Physical weakness on one side" were recorded as different terms: "Left upper limb weakness", " Weakness in the lower left limb " and "Weakness in the lower right limb", which required researchers to manually identify and convert to a local standard term. Certain ways of clinical practices may result in data collection bias, particularly systemic bias. For example, the selected TIA patient group had much higher chance to exhibit various neurological disorders and brain diseases, which were excluded in our risk factor analysis to avoid false result due to potential data bias. In addition, most data from EMRs without coding standards are in unstructured forms, making data standardization difficult. These variations in data can influence CDR calculations. Extra care should be exercised when examining health factors with a high CDR but a limited number of patient node connections. The greater the number of connections, the more reliable the CDR becomes. Once data bias in specific patient populations is identified and recorded, the impacted health factors should be omitted from CDR analysis.

5. Conclusions

This study has generated the first quantitative distribution of TIA risk factors by calculating and sorting the connection delta ratio for each health factor in a hospital's EMR patient graphs. This distribution revealed a potential new risk factor "rhinosinusitis" for future validation. The patient graph was integrated with the standard UMLS TIA knowledge graph, enabling clinicians to search and visualize patient data with standard knowledge in the same graph. The quantification of TIA risk factors by CDR may be applied in development of ML models for TIA risk prediction, risk-based TIA screening and early detection, and TIA management.

Data availability statement

Patient datasets are not available to ensure patient data privacy. Confidential patient data is not deposited in any publicly available repository. Other data without privacy concerns can be obtained from the corresponding author upon a reasonable request.

CRediT authorship contribution statement

Jian Wen: Project administration, Resources, Supervision. **Tianmei Zhang:** Data curation, Writing - review & editing. **Shangrong Ye:** Data curation. **Peng Zhang:** Data curation. **Ruobing Han:** Data curation. **Xiaowang Chen:** Resources. **Ran Huang:** Formal analysis, Software. **Anjun Chen:** Conceptualization, Funding acquisition, Investigation, Methodology, Writing - original draft, Writing - review & editing. **Qinghua Li:** Funding acquisition, Resources, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by Guilin Municipal Science and Technology Bureau, China [grant number 20190219-2], and Sichuan Provincial Science and Technology Bureau, China [grant number 2020YFQ0019]. The authors thank the Guangxi Key Medical and Health Discipline Cultivation and Construction Project for its support. The authors thank Ms. Roufeng Lu for assisting with data and Ms. Erman Wu for editing the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2023.e22766.

References

[1] GBD 2019 Stroke Collaborators. Global, regional, and national burden of stroke and its risk factors, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019, Lancet Neurol. 20 (10) (2021) 795–820, https://doi.org/10.1016/S1474-4422(21)00252-0.
[2] C.W. Tsao, A.W. Aday, Z.I. Almarzooq, et al., Heart disease and stroke statistics—2022 update: a report from the American Heart associationexternal icon, Circulation 145 (8) (2022) e153–e639.
[3] V. Lioutas, C.S. Ivan, J.J. Himali, et al., Incidence of transient ischemic attack and association with long-term risk of stroke, JAMA 325 (4) (2021) 373–381, https://doi.org/10.1001/jama.2020.25071.
[4] P. Amarenco, Steering committee and investigators of the TIAregistry.org Project. Five-year risk of stroke after TIA or minor ischemic stroke, N. Engl. J. Med. 379 (16) (2018) 1580–1581, https://doi.org/10.1056/NEJMc1808913.
[5] D. Kleindorfer, P. Panagos, A. Pancioli, et al., Incidence and short-term prognosis of transient ischemic attack in a population-based study, Stroke 36 (2005) 720–723, https://doi.org/10.1161/01.STR.0000158917.59233.b7.
[6] S. Shahjouei, J. Li, E. Koza, V. Abedi, A.V. Sadr, Q. Chen, A. Mowla, P. Griffin, A. Ranta, R. Zand, Risk of subsequent stroke among patients receiving outpatient vs inpatient care for transient ischemic attack: a systematic review and meta-analysis, JAMA Netw. Open 5 (1) (2022 Jan 4), e2136644, 10.1001/.
[7] Y. Wang, X. Zhao, Y. Jiang, et al., Prevalence, knowledge, and treatment of transient ischemic attacks in China, Neurology 84 (23) (2015) 2354–2361, https://doi.org/10.1212/WNL.0000000000001665.
[8] M.J. Stampfer, P.M. Ridker, V.J. Dzau, Risk factor criteria, Circulation 109 (25 Suppl 1) (2004) IV3–5, https://doi.org/10.1161/01.CIR.0000133446.69171.7d.
[9] R. Ji, L.H. Schwamm, M.A. Pervez, A.B. Singhal, Ischemic stroke and transient ischemic attack in young adults: risk factors, diagnostic yield, neuroimaging, and thrombolysis, JAMA Neurol. 70 (1) (2013) 51–57, https://doi.org/10.1001/jamaneurol.2013.575.
[10] A.W.M. Janssen, F.E. de Leeuw, M.C.H. Janssen, Risk factors for ischemic stroke and transient ischemic attack in patients under age 50, J. Thromb. Thrombolysis 31 (2011) 85–91, https://doi.org/10.1007/s11239-010-0491-3.
[11] M. Schlosser, S. Hazelett, M. Gareri, K. Wright, K. Allen, Incidence of sinusitis in acute ischemic stroke patients, J. Stroke Cerebrovasc. Dis. 12 (5) (2003) 248, https://doi.org/10.1016/j.jstrokecerebrovasdis.2003.11.009.
[12] M. Perez Barreto, S. Sahai, S. Ameriso, J. Ahmadi, D. Rice, M. Fisher, Sinusitis and carotid artery stroke, Ann. Otol. Rhinol. Laryngol. 109 (2) (2000) 227–230, https://doi.org/10.1177/000348940010900220.
[13] K.S. Young, J.S. Kiam, K. Metcalf, et al., Sphenoid sinusitis: a rare cause of ischaemic stroke, BMJ Case Reports CP 14 (2021), e242943.
[14] W.H. Lee, J.W. Kim, J.S. Lim, I.G. Kong, H.G. Choi, Chronic rhinosinusitis increases the risk of hemorrhagic and ischemic stroke: a longitudinal follow-up study using a national sample cohort, PLoS One 13 (3) (2018), e0193886, https://doi.org/10.1371/journal.pone.0193886.
[15] C.W. Wu, P.Z. Chao, W.R. Hao, T.H. Liou, H.W. Lin, Risk of stroke among patients with rhinosinusitis: a population-based study in Taiwan, Am J Rhinol Allergy 26 (4) (2012) 278–282, https://doi.org/10.2500/ajra.2012.26.3783.
[16] S. Zhang, S. Xu, L. Tan, H. Wang, J. Meng, Stroke lesion detection and analysis in MRI images based on deep learning, Journal of Healthcare Engineering (2021), 5524769, https://doi.org/10.1155/2021/5524769.
[17] The Wellcome Trust Case Control Consortium, Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls, Nature 447 (2007) 661–678, https://doi.org/10.1038/nature05911.
[18] T.A. Pearson, T.A. Manolio, How to interpret a genome-wide association study, JAMA 299 (11) (2008) 1335–1344, https://doi.org/10.1001/jama.299.11.1335.
[19] Q. Yuan, T. Cai, C. Hong, et al., Performance of a machine learning algorithm using electronic health record data to identify and estimate survival in a longitudinal cohort of patients with lung cancer, JAMA Netw. Open 4 (7) (2021), e2114723, https://doi.org/10.1001/jamanetworkopen.2021.14723.
[20] S. Tang, P. Davarmanesh, Y. Song, D. Koutra, M.W. Sjoding, J. Wiens, Democratizing EHR analyses with FIDDLE: a flexible data-driven preprocessing pipeline for structured clinical data, J Am Med Inform Assoc 27 (12) (2020) 1921–1934, https://doi.org/10.1093/jamia/ocaa139.
[21] A. Chen, A novel graph methodology for analyzing disease risk factor distribution using synthetic patient data, Healthcare Analytics 2 (2022), 100084, https://doi.org/10.1016/j.health.2022.100084.
[22] O. Bodenreider, The unified Medical Language System (UMLS): integrating biomedical terminology, Nucleic Acids Res. 32 (2004) D267–D270.
[23] S. Timón-Reina, M. Rincón, R. Martínez-Tomás, An overview of graph databases and their applications in the biomedical domain, Database (2021), baab026, https://doi.org/10.1093/database/baab026.
[24] A. Lysenko, I.A. Roznovăţ, M. Saqi, et al., Representing and querying disease networks using graph databases, BioData Min. 9 (2016) 23, https://doi.org/10.1186/s13040-016-0102-8.
[25] D. Martinez, A. Otegi, A. Soroa, E. Agirre, Improving search over Electronic Health Records using UMLS-based query expansion through random walks, J Biomed Inform 51 (2014) 100–106, https://doi.org/10.1016/j.jbi.2014.04.013.
[26] M. Rotmensch, Y. Halpern, A. Tlimat, et al., Learning a health knowledge graph from electronic medical records, Sci. Rep. 7 (2017) 5994, https://doi.org/10.1038/s41598-017-05778-z.
[27] J. Schrodt, A. Dudchenko, P. Knaup-Gregori, M. Ganzinger, Graph-representation of patient data: a systematic literature review, J. Med. Syst. 44 (4) (2020) 86.
[28] J.M. Rogers, J. Bechara, S. Middleton, S.J. Johnstone, Acute EEG patterns associated with transient ischemic attack, Clin. EEG Neurosci. 50 (3) (2019) 196–204, https://doi.org/10.1177/1550059418790708.
[29] U.G.R. Schulz, P.M. Rothwell, Transient ischaemic attacks mimicking focal motor seizures, Postgrad. Med. 78 (2002) 246–247, https://doi.org/10.1136/pmj.78.918.246.
[30] Y.-S. Huang, M. Koo, J.-C. Chen, J.-H. Hwang, The association between tinnitus and the risk of ischemic cerebrovascular disease in young and middle-aged patients: a secondary case-control analysis of a nationwide, population-based health claims database, PLoS One 12 (11) (2017), e0187474, https://doi.org/10.1371/journal.pone.0187474.
[31] F. Hafeez, R.L. Levine, D.A. Dulli, Pulsatile tinnitus in cerebrovascular arterial diseases, J. Stroke Cerebrovasc. Dis. 8 (4) (1999) 217–223, https://doi.org/10.1016/s1052-3057(99)80070-6.
[32] T.N. Turan, J.H. Voeks, M.I. Chimowitz, et al., Rationale, design, and implementation of intensive risk factor treatment in the CREST2 trial, Stroke 51 (10) (2020) 2960–2971, https://doi.org/10.1161/STROKEAHA.120.030730.
[33] V. Abedi, N. Goyal, G. Tsivgoulis, et al., Novel screening tool for stroke using artificial neural network, Stroke 48 (6) (2017) 1678–1681, https://doi.org/10.1161/STROKEAHA.117.017033.
[34] S.F. Weng, J. Reps, J. Kai, J.M. Garibaldi, N. Qureshi, Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS One 12 (4) (2017), e0174944 https://doi.org/10.1371/journal.pone.0174944.
[35] S. Bacchi, L. Oakden-Rayner, T. Zerner, T. Kleinig, S. Patel, J. Jannes, Deep learning natural language processing successfully predicts the cerebrovascular cause of transient ischemic attack-like presentations, Stroke 50 (3) (2019) 758–760, https://doi.org/10.1161/STROKEAHA.118.024124.