Inference of Dependency Knowledge Graph for Electronic Health Records

Zhiwei Xu[1], Ziming Gan[2], Doudou Zhou[3], Shuting Shen[4], Junwei Lu[3*], Tianxi Cai[3,5*]

[1]Department of Statistics, University of Michigan

[2]Department of Statistics, University of Chicago

[3]Department of Biostatistics, Harvard T.H. Chan School of Public Health

[4]Department of Biostatistics and Bioinformatics, Duke University

[5]Department of Biomedical Informatics, Harvard Medical School

Abstract

The effective analysis of high-dimensional Electronic Health Record (EHR) data, with substantial potential for healthcare research, presents notable methodological challenges. Employing predictive modeling guided by a knowledge graph (KG), which enables efficient feature selection, can enhance both statistical efficiency and interpretability. While various methods have emerged for constructing KGs, existing techniques often lack statistical certainty concerning the presence of links between entities, especially in scenarios where the utilization of patient-level EHR data is limited due to privacy concerns. In this paper, we propose the first inferential framework for deriving a sparse KG with statistical guarantee based on the dynamic log-linear topic model proposed by Arora et al. (2016). Within this model, the KG embeddings are estimated by performing singular value decomposition on the empirical pointwise mutual information matrix, offering a scalable solution. We then establish entrywise asymptotic normality for the KG low-rank estimator, enabling the recovery of sparse graph edges with controlled type I error. Our work uniquely addresses the under-explored domain of statistical inference about non-linear statistics under the low-rank temporal dependent models, a critical gap in existing research. We validate our approach through extensive simulation studies and then apply the method to real-world EHR data in constructing clinical KGs and generating clinical feature embeddings.

Keywords: low-rank models, non-linear structure, knowledge graph embedding, hypothesis testing

*Lu and Cai contributed equally

1 Introduction

Electronic Health Record (EHR) data is increasingly recognized for its potential to revolutionize healthcare research (Ahuja et al., 2022; Wen et al., 2023). Its high-dimensional nature, however, poses significant methodological challenges. Constructing clinical knowledge graphs (KGs) from EHR data has become a prevalent approach, offering an effective way to understand the complex interrelations among diverse EHR features. This understanding is crucial for enhancing the efficiency of predictive modeling tasks in various medical applications, such as drug analysis and disease diagnosis (Sang et al., 2018; Abdelaziz et al., 2017). The utility of these graphs extends to improving clinical decision-making (Bauer-Mehren et al., 2013; Finlayson et al., 2014; Nelson et al., 2022) and facilitating the integration and sharing of EHR data (Hong et al., 2021; Abu-Salih et al., 2023).

In recent years, a range of methods for constructing KGs has emerged, notably in the form of KG embeddings. These methods, including translation-based models (Bordes et al., 2013; Wang et al., 2014, e.g.), tensor factorization-based models (Nickel et al., 2011; Yang et al., 2014, e.g.), and neural network-based models (Socher et al., 2013; Bordes et al., 2014, e.g.), effectively map entities and relations into a low-dimensional vector space, capturing the semantics and structure of KGs. Following the advent of word embedding algorithms in natural language processing (Mikolov et al., 2013), these KG embeddings utilize distributed representation technology and address data sparsity and computational inefficiency challenges (Dai et al., 2020). However, a critical limitation of existing techniques is the lack of statistical certainty in the presence of links between entities. This uncertainty quantification is particularly crucial in healthcare applications where accurate and reliable data interpretations are vital for patient

care and treatment decisions. Most KG algorithms in EHR data analysis (Rotmensch et al., 2017; Chen et al., 2019; Zhang et al., 2020; Shang et al., 2021; Harnoune et al., 2021; Roy and Pan, 2021; Jiang et al., 2022), however, can not provide this level of certainty, especially when the use of patient-level data is constrained by privacy concerns.

To fill this gap, we propose the first inferential framework for deriving a sparse KG with uncertainty quantification. This framework is based on the dynamic log-linear topic model proposed by Arora et al. (2016). Under this model, the KG embeddings of the EHR entities can be efficiently approximated via a low-rank representation of the population pointwise mutual information (PMI) matrix. As such, these embeddings can be estimated by performing a singular value decomposition (SVD) of the empirical PMI matrix. We derive the asymptotic properties of this estimator to address estimation uncertainty and provide statistical guidance for edge selection in KGs. To estimate the variance for the estimator, we propose two methods. The first leverages patient-level co-occurrence data for precise variance estimation. While this method offers high accuracy, it raises concerns about data privacy and requires substantial computational resources, particularly with large datasets. We hence propose the second method for situations where patient-level data is not available. Here, we approximate the variance of the estimator under a global null hypothesis of no dependency between EHR entities. This method, prioritizing patient privacy, involves aggregating and summarizing patient data before analysis. It offers improved computational efficiency and enhanced privacy protection. This alternative has shown considerable promise in EHR data applications, balancing the need for privacy with computational efficiency. Leveraging the asymptotic normality of our results, we conduct hypothesis testing on the existence of the edges and incorporate the Benjamini-Hochberg (BH) procedure under dependence (Benjamini

and Yekutieli, 2001) to regulate the False Discovery Rate (FDR).

1.1 Our Contributions

The dynamic log-linear topic model, foundational to our work, has been adapted in several subsequent studies (Arora et al., 2018; Lu et al., 2023; Zhou et al., 2023). However, these adaptations have primarily focused on estimation, leaving a gap in uncertainty quantification. Undertaking inferential analysis on KGs presents three major challenges. For data generation, the model's hierarchical structure entangles several time series and brings the data with a complex dependency structure. For estimation, two layers of non-linearity brought about by the logarithmic and division operators shown later in (2.4) further complicate the analysis. For data privacy, in many situations, we do not have access to patient-level data thus causing the estimation of covariance structure a problem.

Our work contributes to the field of statistical inference in low-rank models, a domain fraught with challenges. Performing statistical inference in such models is particularly difficult due to the intricate nature of low-rank structures and the complexity of their estimation processes. Recent advances have been made in linear low-rank models, including matrix completion (Foucart et al., 2017; Chen et al., 2019), principal component analysis (Xia, 2021; Xia et al., 2022) and low-rank matrix regression (Carpentier et al., 2019; Chernozhukov et al., 2023). However, extending these inferential techniques to non-linear pointwise mutual information under temporal dependent models remains a formidable challenge.

Our research makes several pivotal contributions. First, we enhance the understanding of the low-rank nature of the population PMI by deriving a shaper approximation rate using

the KG embeddings. In specific, instead of assuming certain prior of the embeddings in the existing works (Arora et al., 2018; Lu et al., 2023), the embeddings in our model are deterministic vectors and thus make hypotheses on the cosine similarities between embeddings well-defined. The sharper rate is also crucial for valid inference. Second, we provide a method for uncertainty quantification of the estimator by demonstrating its asymptotic normality, addressing a significant gap where no previous valid inference method for non-linear low-rank models. The complexity of this analysis stems from the high non-linearity of the generative model and the dependencies inherent in longitudinal EHR datasets. Lastly, our study introduces an inference procedure for the KG using only summary-level EHR data. This approach is especially relevant when patient-level data sharing is not feasible, allowing for effective analysis that adheres to privacy constraints.

1.2 Paper Organization

The rest of the paper is organized as follows. Section 2 introduces the generative model of EHR data and proposes the method for KG inference. Section 3 provides the theoretical properties of the empirical PMI matrix and the entry-wise asymptotic normality of the low-rank estimator. Section 4 shows the simulation results and Section 5 applies the proposed method to construct a KG of four domains of EHR entities based on the empirical PMI matrix along with other summary statistics derived from EHR data of over 12 million patients in the Veteran Affairs (VA) health system. The conclusion is given in Section 6. In addition, the proof of the theorems is developed in the supplementary.

2 Method

2.1 Notation

We use $\mathbf{1}_p$ to represent a p-dimensional vector with all entries equal to 1 and $\mathbf{I}_p$ to represent the $p \times p$ identity matrix. Let $\mathcal{O}^{p\times p} = \{\mathbf{R} \in \mathbb{R}^{p\times p} : \mathbf{R}^\top \mathbf{R} = \mathbf{I}_p\}$ denote the set of all $p \times p$ orthonormal matrix. For a vector $\boldsymbol{a} \in \mathbb{R}^p$, we use $\|\boldsymbol{a}\| = \sqrt{\sum_{i=1}^p a_i^2}$ to represent its l_2 norm. For a matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$, we denote by $\mathbf{A}_{i,\cdot} = (A_{i1}, \cdots, A_{im})$ (respectively $\mathbf{A}_{\cdot,i}$) its ith row (respectively column), $\mathbf{A}_{i,i'}$ its (i, i')th element, $\|\mathbf{A}\|_{\max} = \max_{i,j} |\mathbf{A}_{ij}|$ its matrix max norm, $\|\mathbf{A}\|_{2,\infty} = \max_i \|\mathbf{A}_{i,\cdot}\|$ its 2-to-∞ norm, and $\|\mathbf{A}\|$ its matrix operator norm. The Hadamard product of matrices $\mathbf{A}$ and $\mathbf{B}$ is denoted by $\mathbf{A} \circ \mathbf{B}$. If a symmetric matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ has eigen-decomposition $\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$, where $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \cdots, \lambda_n)$, and $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|$ (sort by magnitude), then we use $\lambda_i(\mathbf{A}) = \lambda_i$ to represent the ith largest absolute eigenvalue of $\mathbf{A}$. For two sequences $\{x_n\}$ and $\{y_n\}$, we say $x_n = O(y_n)$ or $x_n \lesssim y_n$ if there exists some constant $C > 0$ such that $x_n \leq C y_n$ for all n. We say $x_n = o(y_n)$ if $\lim_{n\to\infty} x_n/y_n = 0$, and $x_n = \Omega(y_n)$ if $y_n = O(x_n)$. Denote $[d]$ as the set $\{1, 2, \cdots, d\}$. Suppose there are a total of n patients and d unique EHR entities. We let $\Phi(\cdot)$ be the cumulative distribution function of the standard normal distribution.

2.2 Generation of the Longitudinal EHR Feature Occurrences

In this section, we introduce the generative model for the observed longitudinal EHR data. We assume the entities in the EHR are generated from the dynamic log-linear topic model (Arora et al., 2016; Lu et al., 2023). Specifically, for the $i \in [n]$th patient, we observe the longitudinal EHR data $\{w_{i,1}, \ldots, w_{i,t}, \ldots, w_{i,T_i}\}$, where T_i is the last observation time, and

$w_{i,t}$ is the EHR entity observed at time t and follows the following multinomial distribution

$$\mathbb{P}(w_{i,t} = w|\boldsymbol{c}_{i,t}) = \frac{\exp(\langle \mathbf{V}_w, \boldsymbol{c}_{i,t}\rangle)}{\sum_{k=1}^{d} \exp(\langle \mathbf{V}_k, \boldsymbol{c}_{i,t}\rangle)}, \text{ for all } w \in [d], \tag{2.1}$$

where $\mathbf{V}_w$ is the KG embedding of the entity w, and $\boldsymbol{c}_{i,t}$ is the latent discourse vector following a slow autoregressive (AR) process:

$$\boldsymbol{r}_{i,t} \overset{\text{i.i.d.}}{\sim} N(0, \mathbf{I}_p/p), \boldsymbol{c}_{i,1} = \boldsymbol{r}_{i,1}, \boldsymbol{c}_{i,t+1} = \sqrt{\alpha}\boldsymbol{c}_{i,t} + \sqrt{1-\alpha}\boldsymbol{r}_{i,t+1}, \quad t \geq 1, \tag{2.2}$$

where p is the dimension of the embedding, $\alpha = 1 - (\log d)/p^2$ is a measure of the walking rate aligning with that in Lu et al. (2023), $\boldsymbol{r}_{i,t}$ is independent of $\{\boldsymbol{c}_{i,k}\}_{k=1}^{t-1}$; and the discourse vectors are independent across patients. This discourse process is stationary under such setting. We consider the sequence $\{\boldsymbol{c}_{i,t}\}_{t=1}^{T_i}$ follows an AR model different from the sphere random walk as in Arora et al. (2016) to incorporate the temporal variance heterogeneity of the discourse.

Denote the KG embedding matrix by $\mathbf{V} = (\mathbf{V}_1, \ldots, \mathbf{V}_d)^\top$. We see from (2.1) that the probability distribution over the entities is invariant to a constant shift of the rows of $\mathbf{V}$, i.e., for any constant vector $\boldsymbol{\mu} \in \mathbb{R}^p$, replacing $\mathbf{V}$ with $\mathbf{V} - \mathbf{1}_d\boldsymbol{\mu}^\top$ does not change the probability of occurrence. In order to have the identifiable embeddings, we assume with out loss of generality that $\mathbf{V}$ is centered, i.e., $\mathbf{V}^\top\boldsymbol{p} = \mathbf{0}$, where $\boldsymbol{p} = (p_1, \cdots, p_d)^\top \in \mathbb{R}^d$ with $p_w = \mathbb{E}_{\boldsymbol{c}_{i,t}\sim N(0,\mathbf{I}_p/p)}[\mathbb{P}(w_{i,t} = w|\boldsymbol{c}_{i,t})]$ being the marginal occurrence probability of feature w over the stationary discourse process. Thus the identifiability condition $\mathbf{V}^\top\boldsymbol{p} = 0$ implies that the feature embeddings are centered at zero under to the distribution of marginal occurrence

probability $\boldsymbol{p}$.

Under this generative model, the KG can be inferred from the embeddings. This is based on the principle that clinically interconnected entities often appear together. For instance, if the latent discourse vector $\boldsymbol{c}_{i,t}$ pertains to Alzheimer's Disease, there is a high probability of observing both the diagnosis code for "dementias" and a prescription for "memantine." This suggests that the knowledge embeddings for these entities are spatially proximate to each other and also near $\boldsymbol{c}_{i,t}$ in the embedding space. More precisely, the determination of whether an edge exists between any two entities w and w' hinges on hypothesis testing as follows:

$$\mathrm{H}_0 : \mathbf{V}_w^\top \mathbf{V}_{w'} = 0 \quad \text{versus} \quad \mathrm{H}_1 : \mathbf{V}_w^\top \mathbf{V}_{w'} \neq 0. \tag{2.3}$$

The utilization of KG embeddings for predicting edges has gained popularity in constructing KGs. For instance, Hong et al. (2021) employed cosine similarity, defined as $\frac{\mathbf{V}_w^\top \mathbf{V}_{w'}}{\|\mathbf{V}_w\| \|\mathbf{V}_{w'}\|}$, to ascertain the edge weight between w and w'. Nickel et al. (2011) introduced the relational learning approach RESCAL, using the score $\mathbf{V}_w^\top \mathbf{R} \mathbf{V}_{w'}$ for weight prediction, where $\mathbf{R} \in \mathbb{R}^{p \times p}$ is a relation matrix to be learned. Yang et al. (2014) later simplified this by constraining $\mathbf{R}$ to be a diagonal matrix. However, these methods lack a mechanism to quantify uncertainty in the existence of an edge between w and w'. They primarily focus on calculating and selecting high-scoring entity pairs, without a clear threshold to differentiate linked from unlinked entities, nor a way to control the false discovery rate in link discovery.

Our approach, encapsulated in the hypothesis testing procedure (2.3), addresses this gap. By applying this hypothesis testing to all entity pairs (w, w') where $w \neq w'$ within the set $[d] \times [d]$, we can construct a sparsely connected KG. This testing is equivalent to assessing

the cosine similarity

$$\mathrm{H}_0 : \frac{\mathbf{V}_w^\top \mathbf{V}_{w'}}{\|\mathbf{V}_w\| \cdot \|\mathbf{V}_{w'}\|} = 0 \quad \text{versus} \quad \mathrm{H}_1 : \frac{\mathbf{V}_w^\top \mathbf{V}_{w'}}{\|\mathbf{V}_w\| \cdot \|\mathbf{V}_{w'}\|} \neq 0\,,$$

thus bridging the gap with methods that use cosine similarities (Beam et al., 2019; Hong et al., 2021; Zhou et al., 2022, e.g.). It also aligns with semantic matching models (Nickel et al., 2011; Yang et al., 2014, e.g.) by treating the relation matrix $\mathbf{R}$ as an identity matrix.

2.3 Inferential Procedure through Low-Rank Approximation

The challenge for testing (2.3) lies in the fact that the embedding matrix $\mathbf{V}$ is not directly observable and must be inferred from the data $\{w_{i,1}, \ldots, w_{i,t}, \ldots, w_{i,T_i}\}_{i=1}^n$. Utilizing likelihood for this purpose is impractical due to the extensive number of parameters and the complex non-linearity of the probability mass function. This issue is further exacerbated when patient-level data are off-limits due to privacy concerns. Fortunately, it has been demonstrated that these embeddings can be effectively recovered from the PMI matrix, offering a viable solution to this problem. Specifically, the PMI between two entities w and w' is defined as

$$\mathbb{PMI}(w, w') = \log \frac{p_{w,w'}}{p_w p_{w'}}, \tag{2.4}$$

where $p_{w,w'} = \sum_{u=1}^q \mathbb{E}_{(\boldsymbol{c}_{i,t}, \boldsymbol{c}_{i,t+u}) \sim F_u} [\mathbb{P}(w_{i,t} = w, w_{i,t+u} = w' | \boldsymbol{c}_{i,t}, \boldsymbol{c}_{i,t+u})]/q$ denotes the co-occurrence probability of entities w and w' within a pre-defined window of size q, and F_u is the stationary distribution of $(\boldsymbol{c}_{i,t}, \boldsymbol{c}_{i,t+u})$. With the generation model (2.1) in place, we can

show the relationship between the KG embeddings and the PMI matrix as follows:

$$\mathbb{PMI} \approx \alpha_p \mathbf{V}\mathbf{V}^\top, \text{ where } \mathbb{PMI} = [\mathbb{PMI}(w, w')]_{w,w' \in [d]} \in \mathbb{R}^{d \times d} \text{ and } \alpha_p = \frac{\sqrt{\alpha}(1 - \alpha^{q/2})}{qp(1 - \sqrt{\alpha})}. \quad (2.5)$$

Here α is the scaling factor of the AR process (2.2) and $\sqrt{\alpha_p}$ is the scaling factor of the embeddings. A more formal argument is presented in Theorem 3.4. The relationship between the PMI matrix and the KG embedding implies that we can derive the KG of the associations among different EHR entities from the magnitude of the PMI matrix entries.

To estimate the PMI matrix, we first define the co-occurrence between the entities w and w' within the window of size q in all patients' health records as

$$\mathbb{C}_{w,w'} = \sum_{i=1}^{n} \mathbb{C}^{(i)}_{w,w'},$$
$$\text{where } \mathbb{C}^{(i)}_{w,w'} = |\{(t, s) \in [T_i] \times [T_i] : 0 < |t - s| \leq q \text{ and } w_{i,t} = w, w_{i,s} = w'\}|.$$

We then denote $\mathbb{C}^{(i)} = [\mathbb{C}^{(i)}_{w,w'}]_{w,w' \in [d]}$ as the ith patient's co-occurrence matrix, $\mathbb{C} = [\mathbb{C}_{w,w'}]_{w,w' \in [d]}$ as the total co-occurrence matrix, $\mathbb{C}^{(i)}_w = \sum_{w'} \mathbb{C}^{(i)}_{w,w'}$ the individual marginal occurrence of w, and $\mathbb{C}_w = \sum_{w'=1}^{d} \mathbb{C}_{w,w'}$ the marginal occurrence of w, and finally, $\overline{\mathbb{C}}^{(i)} = \sum_w \mathbb{C}^{(i)}_w$ and $\overline{\mathbb{C}} = \sum_w \mathbb{C}_w$.

An empirical estimation for the occurrence probabilities and co-occurrence probabilities is

$$\widehat{p}_w = \mathbb{C}_w / \overline{\mathbb{C}} \text{ and } \widehat{p}_{w,w'} = \mathbb{C}_{w,w'} / \overline{\mathbb{C}}, \quad \text{for } 1 \leq w, w' \leq d. \quad (2.6)$$

We plug in these estimations in the definition of PMI matrix (2.4), and obtain an empirical estimator $\widehat{\mathbb{PMI}} = [\widehat{\mathbb{PMI}}(w, w')]_{w,w' \in [d]}$ for the population PMI matrix $\mathbb{PMI}$ based on the

co-occurrence matrix $\mathbb{C}$, i.e.,

$$\widehat{\mathbb{PMI}}(w, w') = \log \frac{\widehat{p}_{w,w'}}{\widehat{p}_w \widehat{p}_{w'}} = \log \frac{\overline{\mathbb{C}} \cdot \mathbb{C}_{w,w'}}{\mathbb{C}_w \mathbb{C}_{w'}}. \tag{2.7}$$

Remark 2.1. Considering that some co-occurrence entries may be equal to zero, we can set a threshold to $\widehat{\mathbb{PMI}}(w, w')$ to guarantee that it is well defined. In particular, we could use $\log(\max\{\overline{\mathbb{C}} \cdot \mathbb{C}_{w,w'}/(\mathbb{C}_w \mathbb{C}_{w'}), \eta\})$ to avoid the singularity case when the co-occurrence is zero. Here $\eta > 0$ is a threshold to be chosen. For the rest of the paper, we only consider the estimator in (2.7) for simplicity by showing the co-occurrence is non-zero with high probability.

While we can directly apply $\widehat{\mathbb{PMI}}$ to do inference on $\mathbf{V}\mathbf{V}^\top$, this method lacks efficiency as it overlooks the low-rank structure of $\mathbf{V}\mathbf{V}^\top$. To utilize the low-rank structure, we propose to use the rank-p approximation of $\widehat{\mathbb{PMI}}$ as the ultimate estimator of $\alpha_p \mathbf{V}\mathbf{V}^\top$. Namely, denote the eigendecomposition of $\widehat{\mathbb{PMI}}$ as

$$\widehat{\mathbb{PMI}} = \begin{bmatrix} \widehat{\mathbf{U}} & \widehat{\mathbf{U}}_\perp \end{bmatrix} \begin{bmatrix} \widehat{\mathbf{\Lambda}} & \mathbf{0} \\ \mathbf{0} & \widehat{\mathbf{\Lambda}}_\perp \end{bmatrix} \begin{bmatrix} \widehat{\mathbf{U}}^\top \\ \widehat{\mathbf{U}}_\perp^\top \end{bmatrix},$$

where $(\widehat{\mathbf{U}}, \widehat{\mathbf{U}}_\perp) \in \mathbb{R}^{d \times d}$ are the eigenvectors with the eigenvalues sorted in descending order by their magnitude and $\widehat{\mathbf{U}} \in \mathbb{R}^{d \times p}$. Then we let

$$\widetilde{\mathbb{PMI}} = \widehat{\mathbf{U}} \widehat{\mathbf{\Lambda}} \widehat{\mathbf{U}}^\top \tag{2.8}$$

be the low-rank estimator of $\alpha_p \mathbf{V}\mathbf{V}^\top$. In the following section, we demonstrate that, under

certain assumptions, $\widetilde{\mathbb{PMII}}$ exhibits entry-wise asymptotic normality and the column space spanned by $\widehat{\mathbf{U}}$ closely aligns with the column space spanned by $\mathbf{V}$. We will show that the low-rank estimator $\widetilde{\mathbb{PMII}}$ possesses lower variance in comparison to the empirical estimator $\widehat{\mathbb{PMII}}$ and minimal bias relative to $\alpha_p \mathbf{V}\mathbf{V}^\top$. This finding suggests that $\widetilde{\mathbb{PMII}}$ is a more efficient estimator, and we will employ it as the foundation for subsequent inferential procedures.

2.3.1 Variance Estimation with Patient-level Data

We now discuss how to estimate the covariance of the estimators using the patient-level co-occurrence matrices $\{\mathbb{C}_i\}_{i=1}^n$. Denote the error matrix by $\mathbf{W} = \widehat{\mathbb{PMII}} - \alpha_p \mathbf{V}\mathbf{V}^\top$. We can express the leading term of the error as $\mathbf{W}_{w,w'} \approx \frac{1}{n}\sum_{i=1}^n \widehat{\mathbf{W}}^{(i)}_{w,w'}$, where

$$\widehat{\mathbf{W}}^{(i)}_{w,w'} = \frac{\mathbb{C}^{(i)}_{w,w'}}{\mathbb{C}_{w,w'}/n} - \frac{\mathbb{C}^{(i)}_{w}}{\mathbb{C}_{w}/n} - \frac{\mathbb{C}^{(i)}_{w'}}{\mathbb{C}_{w'}/n} + 1, \quad 1 \le i \le n.$$

The detail of this approximation is given in Supplementary S2.1. Therefore, we can estimate the covariance between $\widehat{\mathbb{PMII}}_{w,\cdot}$ and $\widehat{\mathbb{PMII}}_{w',\cdot}$, i.e., $\mathrm{Cov}(\widehat{\mathbb{PMII}}_{w,\cdot}, \widehat{\mathbb{PMII}}_{w',\cdot})$ by

$$\widehat{\mathbf{\Sigma}}_{w,w'} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \widehat{\mathbf{W}}^{(i)\top}_{w,\cdot} \widehat{\mathbf{W}}^{(i)}_{w',\cdot}. \tag{2.9}$$

Then the variance of the low-rank estimator $\widetilde{\mathbb{PMII}}$ can be estimated by

$$\widehat{\mathrm{Var}}(\widetilde{\mathbb{PMII}}(w,w')) = \widehat{\mathbf{P}}_{w,\cdot}\widehat{\mathbf{\Sigma}}_{w',w'}\widehat{\mathbf{P}}^\top_{w,\cdot} + \widehat{\mathbf{P}}_{w',\cdot}\widehat{\mathbf{\Sigma}}_{w,w}\widehat{\mathbf{P}}^\top_{w',\cdot} + 2\widehat{\mathbf{P}}_{w,\cdot}\widehat{\mathbf{\Sigma}}_{w',w}\widehat{\mathbf{P}}^\top_{w',\cdot} \tag{2.10}$$

where $\widehat{\mathbf{P}} = \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top$. We refer to Theorem 3.6 for more details. However, due to data privacy, we sometimes have no access to each patient's co-occurrence information, which prevents us

from estimating the covariance via (2.9). In the following section, we will propose a summary statistic-based covariance estimator.

2.3.2 Variance Estimation without Patient-level Data

If we cannot access the patient-level data, we can still test the hypothesis on the significance of $\mathbf{V}_w^\top \mathbf{V}_{w'}$ by testing the global null hypothesis that the occurrence of all EHR entities are independent, i.e.,

$$H_0 : w_{i,t} \overset{i.i.d}{\sim} \text{Multinomial}(1, \{p_w\}_{w=1}^d) \text{ for } 1 \le t \le T_i, 1 \le i \le n.$$

Under the global null distribution, we can estimate the covariance structure of $\mathbf{W}$ without the need for patient-level data using the following formulas (see details in Supplementary S2.5):

$$\begin{aligned}
\widehat{\text{Cov}}\big(\mathbf{W}_{w\cdot}\big) &= \frac{1}{nT_0\widehat{p}_w}\Big(\mathbf{1}\mathbf{1}^\top(\widehat{p}_w - \frac{1}{2}) - \frac{1}{2}\mathbf{1}\boldsymbol{e}_w^\top - \frac{1}{2}\boldsymbol{e}_w\mathbf{1}^\top + \text{diag}(\widehat{\boldsymbol{p}})^{-1}\frac{1-\widehat{p}_w}{2} + \frac{1}{2\widehat{p}_w}\boldsymbol{e}_w\boldsymbol{e}_w^\top\Big), \\
\widehat{\text{Cov}}\big(\mathbf{W}_{w\cdot}, \mathbf{W}_{w'\cdot}\big) &= \frac{1}{nT_0}\Big(\mathbf{1}\mathbf{1}^\top - \frac{1}{2\widehat{p}_w}\mathbf{1}\boldsymbol{e}_w^\top - \frac{1}{2\widehat{p}_{w'}}\boldsymbol{e}_{w'}\mathbf{1}^\top - \frac{1}{2}\text{diag}(\widehat{\boldsymbol{p}})^{-1} + \frac{1}{2\widehat{p}_w\widehat{p}_{w'}}\boldsymbol{e}_{w'}\boldsymbol{e}_w^\top\Big),
\end{aligned} \tag{2.11}$$

where $\widehat{\boldsymbol{p}} = (\widehat{p}_1, \cdots, \widehat{p}_d)^\top$, $T_0 = Tq - q(q+1)/2$ and $T = \sum_{i=1}^n T_i/n$. To estimate the covariance matrices of $\mathbf{W}$, we only need to know parameters n, T, q, and the estimate of $\{p_w\}_{w=1}^d$, which can be obtained from the summary-level co-occurrence matrix $\mathbb{C}$ by (2.6). With $\widehat{\text{Cov}}(\mathbf{W}_{w,\cdot}, \mathbf{W}_{w',\cdot})$ obtained, we can further estimate the variance of $\widetilde{\mathbb{PMI}}(w, w')$ by

(2.10) with $\widehat{\mathbf{\Sigma}}_{w,w'}$ replaced by $\breve{\mathbf{\Sigma}}_{w,w'} = \widehat{\mathrm{Cov}}\big(\mathbf{W}_{w\cdot}, \mathbf{W}_{w'\cdot}\big)$ as

$$\widetilde{\mathrm{Var}}(\widetilde{\mathbb{PMI}}(w,w')) = \widehat{\mathbf{P}}_{w,\cdot}\breve{\mathbf{\Sigma}}_{w',w'}\widehat{\mathbf{P}}_{w,\cdot}^{\top} + \widehat{\mathbf{P}}_{w',\cdot}\breve{\mathbf{\Sigma}}_{w,w}\widehat{\mathbf{P}}_{w',\cdot}^{\top} + 2\widehat{\mathbf{P}}_{w,\cdot}\breve{\mathbf{\Sigma}}_{w',w}\widehat{\mathbf{P}}_{w',\cdot}^{\top}, \tag{2.12}$$

and conduct the hypothesis testing based on the asymptotic normality of $\widetilde{\mathbb{PMI}}$. The KNowledge graph Inference and Testing (KNIT) procedure is then summarized in Algorithm 1.

Algorithm 1 KNowledge graph Inference and Testing (KNIT)

Input: the summary-level co-occurrence matrix $\mathbb{C} \in \mathbb{R}^{d\times d}$, the window size q, the number of patients n, the rank of KG embeddings r, the average observation time $T = \sum_{i=1}^{n} T_i/n$, and the nominal FDR level α.
Estimation: Calculate $\widehat{\mathbb{PMI}}$ and $\widetilde{\mathbb{PMI}}$ based on $\mathbb{C}$ using (2.7) and (2.8).
Variance estimation: Obtain $\widetilde{\mathrm{Var}}(\widetilde{\mathbb{PMI}}(w,w'))$ by (2.12) for $1 \le w < w' \le d$.
Inference: Obtain the p-value $\alpha_{w,w'} = \Phi\big(\widetilde{\mathbb{PMI}}(w,w')/\sqrt{\widetilde{\mathrm{Var}}(\widetilde{\mathbb{PMI}}(w,w'))}\big)$ for $1 \le w < w' \le d$. Order $\alpha_{w,w'}$'s as $\alpha_{(1)} \le \alpha_{(1)} \le \ldots \le \alpha_{(J)}$ with $J = d(d-1)/2$ and set $\alpha_{(0)} = 0$. Let $j_{\max} = \max\{0 \le j \le J : \alpha_{(j)} \le \alpha j/J/(\log J + 1)\}$.
Output: $\widetilde{\mathbb{PMI}}$ and the edge set $\{(w,w') : \alpha_{w,w'} \le \alpha_{j_{\max}}\}$.

3 Theoretical Properties

In this section, we establish the inferential results by characterizing the asymptotic distribution of $\widetilde{\mathbb{PMI}}$. Given our data generative model, only the left singular space and the singular values of $\mathbf{V}$ are identifiable, since for any orthonormal matrix $\mathbf{O} \in \mathbb{R}^{p\times p}$, $\{\boldsymbol{c}_t\}$ and $\{\mathbf{O}\boldsymbol{c}_t\}$ have the same distribution, and replacing $\mathbf{V}$ with $\mathbf{VO}$ does not change the occurrence probabilities in (2.1). Hence without loss of generality, we assume $\mathbf{V}$ has the singular value decomposition $\mathbf{V} = \mathbf{U}\mathbf{\Lambda}^{1/2}$. We denote $\mathbf{\Lambda}^{1/2} = \mathrm{diag}(\kappa_1/\sqrt{\alpha_p}, \cdots, \kappa_p/\sqrt{\alpha_p})$ with $\kappa_1/\sqrt{\alpha_p} \ge \ldots \ge \kappa_p/\sqrt{\alpha_p} > 0$ being the corresponding singular values. Denote $\kappa := \max_{1\le j\le p}|\kappa_j|, \xi := \min_j |\kappa_j|/\kappa$. We require the following assumptions on these scaling parameters.

Assumption 3.1 (Scaling)**.** Assume that there exist constants $0 < c < C$ such that $p \geq c \log^2 d$ and $\log p \leq C \log(\log d)$. We suppose $T_i = T$ for each i, and we have $cnd \log^2 d \leq T \leq Cd^{4 \log d}$ and $\log d / \log n = o(1)$.

Remark 3.1. Assumption 3.1 states the necessary scaling conditions to guarantee that the estimator converges to the truth at a fast enough rate. The lower bound for p is to control the potential information loss caused by low-rank approximation, and the upper bound for p is to control the variance of the estimator $\widetilde{\mathbb{PMII}}$ and maintain the "near low-rank" structure of the population PMI matrix. Although we could consider $\{T_i\}_{i=1}^n$ to be of the same order, leading to $T_i/T_j = O(1)$ for any $i, j \in [n]$, we opt for the simpler assumption of $T_i = T$ for the ease of expression. Here we impose the upper bound for T because when T grows exceptionally large, the advantages of employing the low-rank estimator $\widetilde{\mathbb{PMII}}$ over its empirical counterpart $\widehat{\mathbb{PMII}}$ disappear. The intricate Markov dependency structure in (2.2) necessitates the scaling assumption on the sample size n, ensuring the signal is significant enough to conduct the inference.

Assumption 3.2. Assume that there exists constants $\mu_1 \geq 1, \mu_2 > 0$ such that $\|\mathbf{U}\|_{2,\infty} \leq \mu_1 \sqrt{p/d}$ and $\min_{i \in [d]} \|\mathbf{U}_{i,\cdot}\| \geq \mu_2 p^2 / \xi d$.

Remark 3.2. The upper bound for $\|\mathbf{U}\|_{2,\infty}$ imposes the incoherence condition (Candès and Tao, 2010) on the eigen-space of $\mathbf{V}$. Concurrently, the lower bound for $\min_i \|\mathbf{U}_{i,\cdot}\|$ is very weak since it is of order $p^2/\xi d$, which decays much faster than the upper bound combined with Assumption 3.3. These two inequalities together guarantee that the eigenvectors are roughly uniformly distributed in the eigen-space.

Assumption 3.3. Assume that $\xi = \Omega(1/\log d)$, $\kappa^2 = \Omega(d^6 \log^4 d(nT)^{-1/2}\xi^{-6})$, and $\kappa^2 = O(dp^{-1/2}(nT)^{-1/4})$.

Remark 3.3. The upper bound for κ arises from the variance-bias tradeoff. We need the bias, i.e., the error between $\mathbb{PMI}$ and its low-rank approximation to be comparable to the variance. The lower bound assumption for κ is necessary for reasons similar to the lower bound for $\min_i \|\mathbf{U}_{i,\cdot}\|$ in Assumption 3.2: an eigen-gap is required to prevent the collapse of the low-rank estimator and to control the distance between two singular spaces according to the Davis-Kahan theorem (Davis and Kahan, 1970). An upper bound for the ratio of singular values ξ ensures that the eigen-gap is relatively large compared to the maximum distance within non-zero singular values.

Under the assumptions above, we can measure the distance between the population PMI matrix and the inner product of the scaled KG embeddings, and show the asymptotic normality of the empirical estimator and the low-rank estimator. The following theorem demonstrates the "near low-rank" structure of the population PMI matrix. Its proof is given in Supplementary S1.3.1.

Theorem 3.4. Suppose Assumptions 3.2 and 3.3 hold. Then we have

$$\|\mathbb{PMI} - \alpha_p \mathbf{V}\mathbf{V}^\top\|_{\max} \leq C\frac{\kappa^4 p^2}{d^2}$$

for some constant $C > 0$. Besides, we have $\lambda_1(\mathbb{PMI}) \leq 2\kappa^2, \lambda_p(\mathbb{PMI}) \geq \kappa^2\xi^2/2$, and $|\lambda_{p+k}(\mathbb{PMI})| \leq C\kappa^4 p^2/d$ for $1 \leq k \leq d-p$.

Remark 3.4. This theorem provides the rate of the bias, which is sharper than the

bound $\|\mathbb{PMI} - \mathbf{V}\mathbf{V}^\top/p\|_{\max} = O(1/p)$ given in Arora et al. (2016) and the bound $\|\mathbb{PMI} - \mathbf{V}\mathbf{V}^\top/p\|_{\max} = O(\sqrt{\log(d)/p})$ given in Lu et al. (2023), both of which are not sufficient for the inference of PMI estimator. In comparison, our result reveals a better low-rank approximation when $p = o(d^{2/3})$ as implied by Assumption 3.1. Moreover, our low-rank approximation differs from Arora et al. (2016); Lu et al. (2023)'s as they assumed a mean-zero Gaussian or spherical Gaussian prior on $\mathbf{V}$, while our study does not impose any prior on $\mathbf{V}$, yielding more generalized results.

With the bias bound above, we have the following lemma on the asymptotic normality of the error term.

Lemma 3.5. Suppose Assumptions 3.1-3.3 hold, then we have the entry-wise asymptotic normality of the residual matrix $\mathbf{W}$, i.e., for any entry $(w, w') \in [d] \times [d]$, we have

$$\frac{\sqrt{n}}{\sigma_{w,w'}}\mathbf{W}_{w,w'} \xrightarrow{d} N(0,1) \text{ as } n, d, T \to \infty, \tag{3.1}$$

where $\sigma^2_{w,w'} = \mathrm{Var}\big(\mathbb{C}^{(i)}_{w,w'}/\mathbb{E}[\mathbb{C}^{(i)}_{w,w'}] - \mathbb{C}^{(i)}_{w}/\mathbb{E}[\mathbb{C}^{(i)}_{w}] - \mathbb{C}^{(i)}_{w'}/\mathbb{E}[\mathbb{C}^{(i)}_{w'}]\big)$ is of order d^2/T.

Lemma 3.5 gives the asymptotic normality of the residual matrix $\mathbf{W}$. The proof of this lemma can be found in Supplementary S2.2. We denote the covariance matrix of the error term in (3.1) as $\mathbf{\Sigma}_E$ where $\mathbf{\Sigma}_E(w, w') = \sigma^2_{w,w'}$ for each $w, w' \in [d]$. The following theorem gives the asymptotic distribution of the low-rank estimator.

Theorem 3.6. Let $\widetilde{\mathbf{\Sigma}} = (\mathbf{P}^\star \circ \mathbf{P}^\star)\mathbf{\Sigma}_E + \mathbf{\Sigma}_E(\mathbf{P}^\star \circ \mathbf{P}^\star)$. Under Assumptions 3.1-3.3, for the

low-rank estimator $\widetilde{\mathbb{PMII}}$, we have for any entry $(w, w') \in [d] \times [d]$,

$$\frac{\sqrt{n}}{\widetilde{\mathbf{\Sigma}}_{w,w'}^{1/2}}(\widetilde{\mathbb{PMII}}_{w,w'} - \alpha_p \mathbf{V}_w^\top \mathbf{V}_{w'}) \xrightarrow{d} N(0,1) \text{ as } n, d, T \to \infty,$$

where $\|\widetilde{\mathbf{\Sigma}}\|_{\max} = O(\frac{dp^2}{T\xi^4})$.

The proof of Theorem 3.6 can be found in Supplementary S2.3. In fact, the proof of Theorem 3.6 provides a Bahadur decomposition of $\widetilde{\mathbb{PMII}} - \alpha_p \mathbf{V}_w^\top \mathbf{V}_{w'}$ to establish the entrywise asymptotic normality for the low-rank estimator $\widetilde{\mathbb{PMII}}$, which can be generalized for multiple testing. Compared with the variance of the empirical estimator $\sigma_{w,w'}^2 = d^2/T$ shown in Lemma 3.5, the variance of the low-rank estimator $\widetilde{\mathbf{\Sigma}}_{w,w'} = O(dp^2/(T\xi^4))$ is much smaller. This suggests the superior performance of the low-rank estimator over the empirical estimator.

4 Simulation

In this section, we conduct synthetic simulations to demonstrate the performance of our algorithm. In Section 4.1, we compare the numerical performance of the low-rank estimator with the empirical estimator. In Section 4.2, we generate data under the global null distribution and calculate type I error for edge selection. It shows that the inference with and without patient-level data both have empirical type I errors close to the nominal significance level. In Section 4.3, we show the asymptotic normality of the standardized estimators.

4.1 Numerical Performance of PMI Estimators

In this section, we compare $\widetilde{\mathbb{PMI}}$ with $\widehat{\mathbb{PMI}}$ in terms of estimation errors. We set the number of entities $d = 100$, embedding dimension $p = \lfloor \log^2 d \rfloor + 1$, $T = 1000$, and the sample size $n \in \{200, 400, 600, 800, 1000, 2000\}$. The construction of the embeddings $\mathbf{V}$ involves three steps:

1. Generate an orthonormal basis of a $d \times p$ matrix, with each element independently drawn from a standard normal distribution. Denote it as $\mathbf{V}^{\text{init}}$.

2. Estimate the marginal occurrence probabilities $\boldsymbol{p}$ using the Monte-Carlo (MC) method with 10^7 i.i.d MC samples. Denote it as $\boldsymbol{p}_{\text{mc}}$.

3. Let $\mathbf{V} = (\mathbf{V}_1, \cdots, \mathbf{V}_d)^\top$, where each $\mathbf{V}_w = \mathbf{V}_w^{\text{init}} - \mathbf{V}^{\text{init}\top}\boldsymbol{p}_{\text{mc}}$ for $w \in [d]$.

Here the last step is to guarantee that $\mathbf{V}^\top \boldsymbol{p} = 0$. Then we generate 100 i.i.d replicates. For each replicate, the sequence $\{w_{i,t}\}_{1 \le i \le n, 1 \le t \le T}$ is generated in accordance with (2.1), and $\|\widehat{\mathbb{PMI}} - \alpha_p \mathbf{V}\mathbf{V}^\top\|_{\max}$, $\|\widetilde{\mathbb{PMI}} - \alpha_p \mathbf{V}\mathbf{V}^\top\|_{\max}$, and $\|\mathbb{PMI} - \alpha_p \mathbf{V}\mathbf{V}^\top\|_{\max}$ are calculated. As evidenced by the first plot in Figure 1, the low-rank estimator has less variability and outperforms the empirical estimator, and the bias $\|\mathbb{PMI} - \alpha_p \mathbf{V}\mathbf{V}^\top\|_{\max}$ is negligible compared to $\|\widehat{\mathbb{PMI}} - \alpha_p \mathbf{V}\mathbf{V}^\top\|_{\max}$ and $\|\widetilde{\mathbb{PMI}} - \alpha_p \mathbf{V}\mathbf{V}^\top\|_{\max}$, which are consistent with our theoretical results.

4.2 Numerical Results on Hypothesis Testing

In this section, we simulate data following the global null distribution given in Section 2.3.2. For simplicity, we assume that each feature occurs with equal probability, such that

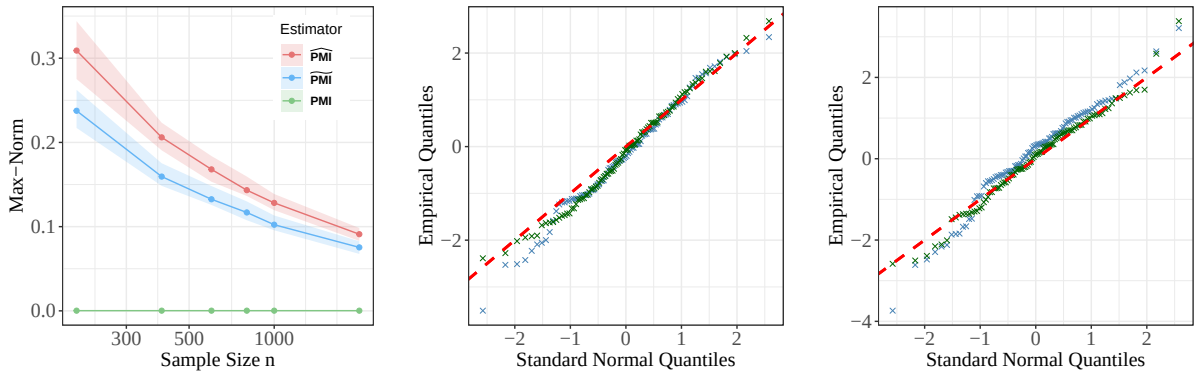

Figure 1: Left plot is the median and the 20%-80% quantile range of max norm distances between word embeddings and PMI estimators across 100 independent runs. Middle and right plots are Q-Q plots of one standardized entry for the empirical (green) estimator and the low-rank (blue) estimator. The Middle: $(d, n, T, \kappa) = (100, 1000, 800, 1), (w, w') = (1, 1)$. Right: $(d, n, T, \kappa) = (100, 1000, 1200, 1), (w, w') = (1, 1)$.

$p_w = 1/d$ for $1 \leq w \leq d$. The sets $\{w_{i,t}\}_{1 \leq i \leq n, 1 \leq t \leq T}$ are independently and identically drawn from a Multinomial$(1, (1/d, \cdots, 1/d))$ distribution, considering $d \in \{100, 200, 300\}$ and $n \in \{200, 400, 800, 1600\}$. Under these conditions, it is not reasonable to define the embeddings, and thus our analysis is confined to the empirical estimator. We apply the methods described in Sections 2.3.1 and 2.3.2 to estimate the variance of each entry in $\widehat{\mathbb{PMI}}$, employing the Monte-Carlo technique to estimate the empirical type I error. This involves generating n i.i.d longitudinal code sequences 100 times, and subsequently calculating the type I errors for each method. When patient-level data is accessible, we estimate the variance of each entry using (2.9). In contrast, when patient-level data is unavailable, we estimate the variance of each entry via (2.11).

The testing is conducted on the ten edges from the 1st entity to the 2nd, 3rd, ..., and 11th entities, with the nominal significance level set as $\alpha = 0.05$ for each edge. We then average these type I errors to estimate the empirical significance level for any arbitrary edge test, with the results presented in Table 1. It is observed that the averages of the ten empirical type I

errors, as calculated by both methods, tend to approach 0.05 when n becomes sufficiently large. This observation underscores the asymptotic normality of the empirical estimator and suggests that even in the absence of patient-level data, it is feasible to estimate the variance of each edge and establish confidence intervals for individual edges.

d	100			200			300			
n	200	400	800	200	400	800	200	400	800	1600
Patient level	0.072	0.061	0.042	0.069	0.067	0.049	0.077	0.077	0.064	0.040
Summary stat.	0.062	0.064	0.048	0.068	0.054	0.050	0.077	0.069	0.062	0.045

Table 1: The averages of the first ten edges' empirical type I errors under the global null distribution using the empirical estimator. The row of "Patient level" uses the patient-level data in Section 2.3.1 and the row of " Summary stat." applies the method in Section 2.3.2 using the total co-occurrence matrix to estimate the variances.

4.3 Numerical Results on the Asymptotic Normality

In this section, we generate data from (2.1). The embeddings are constructed using $\mathbf{V} = d^{-\kappa}\mathbf{U}$, where $\kappa \in \{1, 2\}$ and $\mathbf{U}$ is formed from the orthonormal bases of a $d \times p$ matrix with each element independently drawn from a standard normal distribution. Notably, in this context, we do not utilize $\mathbf{V}\mathbf{V}^\top$, and there is no distinction between using $\mathbf{V}$ and $\mathbf{V} - \mathbf{1}\mu^\top$ for data generation. Consequently, the 2nd and 3rd steps outlined in Section 4.1 are unnecessary for embedding construction.

The second and third plots in Figure 1 depict the QQ plots of both the empirical estimator and the low-rank estimator. They suggest that both estimators are asymptotically normal. In Figure 2, we illustrate that the confidence intervals of $\widetilde{\mathbb{PMI}}$ are consistently narrower than those of $\widehat{\mathbb{PMI}}$. This indicates that the low-rank estimator $\widetilde{\mathbb{PMI}}$ outperforms the empirical estimator which also validates our theoretical analysis on the order of variances.

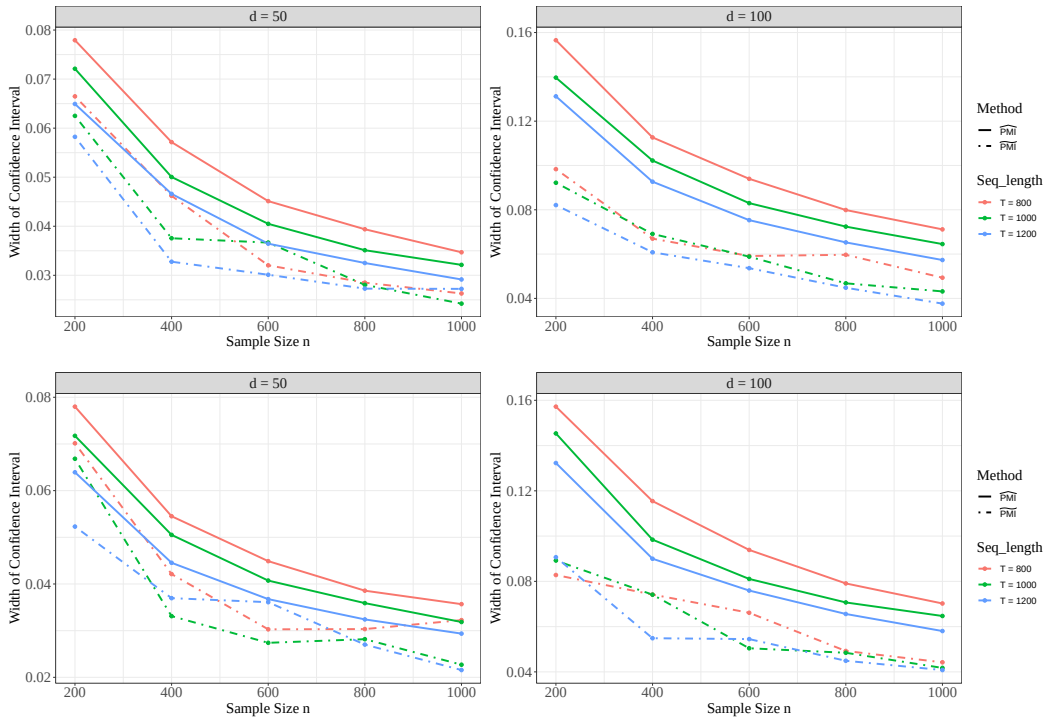

Figure 2: Plots of the average half width of 95% confidence interval of the first ten entries respectively constructed by the empirical estimator and the low-rank estimator. The first row is for $\kappa = 1$ and the second row is for $\kappa = 2$.

5 Applications to Electronic Health Records

We validate the efficacy of our algorithm using EHR data from the Veterans Affairs (VA) Healthcare System consisting of 12.6 million patients who had at least one visit between 2000 and 2019. The dataset includes four specific domains of codified information: PheCodes for diseases, Current Procedural Terminology (CPT) codes for procedures, lab test codes, and RxNorm codes for medication prescriptions. Codes occurring fewer than 5000 times were excluded to minimize noise from infrequent codes, resulting in a final count of 1776 PheCodes, 224 CCS codes, 6025 lab codes, and 1561 RxNorm codes as detailed in Hong et al. (2021). These codes were then utilized to construct a co-occurrence matrix, adopting a window size, q, of 30 days following the analysis of Hong et al. (2021).

To assess our method's performance, we employ known similar and related pairs of these codes as established connections in the KG. Similar code pairs denote codes that embody highly analogous clinical entities according to existing ontologies. In contrast, related code pairs represent codes with more intricate relationships, such as 'may cause' or 'may treat'. For similarity assessments, we leverage the hierarchies of several ontologies to define similar pairs. These include disease-disease pairs from the PheCode Hierarchy and drug-drug pairs from the ATC classification system within the Unified Medical Language System (McInnes et al., 2007). We identified related pairs from online knowledge sources, categorizing them into four types: disease-disease, disease-drug, disease-procedure, and disease-lab. The counts of these pairings are detailed in Table 2, with details provided in Zhou et al. (2022).

Relation Type	Relation Category	Code pairs Number
Similarity	PheCode Hierarchy (Disease-Disease)	4094
	RxNorm-RxNorm (Drug-Drug)	3647
Relatednedd	PheCode-PheCode (Disease-Disease)	2430
	PheCode-RxNorm (Disease-Drug)	5523
	PheCode-CCS (Disease-Procedure)	2545
	PheCode-Lab (Disease-Lab)	334

Table 2: Number of curated relationship pairs categorized by seven categories.

We begin by evaluating the quality of the estimators. In each known-relation pair category, an equal number of negative pairs are randomly sampled. The predicted scores are then determined using our low-rank estimator $\widetilde{\mathbb{PMI}}(w, w')$, compared with the empirical estimator $\widehat{\mathbb{PMI}}(w, w')$. Following this, the Area Under the Receiver Operating Characteristic Curve (AUC) is calculated for both methods, utilizing the positive and negative pairs.

Furthermore, we conduct a comparison with neural network-based embeddings, specifically those generated by various Bidirectional Encoder Representations from Transformer

(BERT)-based algorithms (Devlin et al., 2019) trained on clinical corpora such as Self-aligning Pretrained BERT (SAPBERT) (Liu et al., 2021), BERT for Biomedical Text Mining (BioBERT) (Lee et al., 2020), and BERT pretrained with PubMed (PubmedBERT) (Gu et al., 2021). These algorithms take code descriptions as inputs and produce embeddings as outputs.

The results, as displayed in Table 3, reveal that $\widetilde{\mathbb{PMI}}$ consistently outperforms the other methods in most of the categories, underlining its capability to effectively capture clinical knowledge from extensive EHR data. This demonstrates that $\widetilde{\mathbb{PMI}}$ is a more precise estimator for representing clinical knowledge derived from massive EHR data.

Relation Type	Category	$\widetilde{\mathbb{PMI}}$	$\widehat{\mathbb{PMI}}$	Sap	PubMed	Bio	Bert
Similarity	PheCode Hierarchy	**0.970**	0.856	0.763	0.616	0.580	0.592
	Drug-Drug	**0.835**	0.757	0.611	0.491	0.520	0.471
	average	**0.906**	0.809	0.691	0.557	0.552	0.535
Relatedness	Disease-Disease	**0.843**	0.826	0.635	0.612	0.573	0.564
	Disease-Drug	**0.823**	0.810	0.604	0.631	0.610	0.597
	Disease-Procedure	0.743	**0.760**	0.635	0.594	0.565	0.531
	Disease-Lab	**0.805**	0.790	0.529	0.488	0.602	0.552
	average	**0.808**	0.801	0.616	0.614	0.591	0.573

Table 3: The AUC of detecting known relation pairs with different methods. Sap denotes SapBert, PubMed denotes PubMedBert, Bio denotes BioBert. The row named 'average' exhibits the results of the average AUC weighted by the number of pairs within each category.

We then implement our testing procedure to identify significant edges, setting the target FDR at 0.05 and 0.1 within each category. To showcase the effectiveness of our proposed algorithm, we compare its power with other benchmark methods. Given that no existing methods provide statistical uncertainty quantification for the association between two entities, we rank the similarities of the pairs and select the top-ranking pairs as related pairs for other methods. The number of selected pairs matches that chosen by our testing procedure. This

approach allows us to contrast the power of our algorithm with the benchmark methods.

From the results presented in Table 4, it is evident that our algorithm demonstrates a high power in detecting known-relation pairs. In particular, for several pair categories, such as PheCode-PheCode pairs from PheCode Hierarchy, and related disease-disease pairs, our algorithm achieves an exceptional power of over 85% while all other methods have power lower than 75%. Furthermore, our method consistently outperforms the selection based on empirical PMI or cosine-similarity from other benchmarks for all pair categories. This underscores the capability of our algorithm to not only establish a threshold for selecting significant edges but also to leverage the variance information of the estimator, surpassing the performance of methods reliant solely on cosine similarity. Similar patterns can be found in Table S1 in Supplementary S3 with target FDR set as 0.1.

Relation Type	Category	KNIT	$\widehat{\mathbb{PMI}}$	Sap	PubMed	Bio	Bert
Similarity	PheCode Hierarchy	**0.917**	0.816	0.723	0.514	0.468	0.478
	Drug-Drug	**0.790**	0.587	0.679	0.732	0.785	0.764
	average	**0.857**	0.708	0.702	0.617	0.617	0.613
Relatedness	Disease-Disease	**0.872**	0.833	0.546	0.505	0.467	0.442
	Disease-Drug	**0.805**	0.715	0.367	0.426	0.383	0.381
	Disease-Procedure	0.752	**0.760**	0.521	0.472	0.369	0.355
	Disease-Lab	**0.677**	0.560	0.394	0.363	0.333	0.312
	average	**0.804**	0.747	0.444	0.453	0.397	0.386

Table 4: Power of detecting known relation pairs with our algorithm compared with other benchmarks, under target FDR being 0.05. Sap denotes SapBert, PubMed denotes PubMedBert, Bio denotes BioBert. The row named 'average' exhibits the results of the average power in detecting similar (related) pairs weighted by the number of pairs within each category.

To showcase the efficacy of our algorithm in constructing a KG for clinical entities, we focus specifically on Alzheimer's Disease (AD) as a case study. We present in Figure 3 the p-values and $\widetilde{\mathbb{PMI}}_{w,w'}$, where w is AD and w''s are the other top $5,000$ entities with

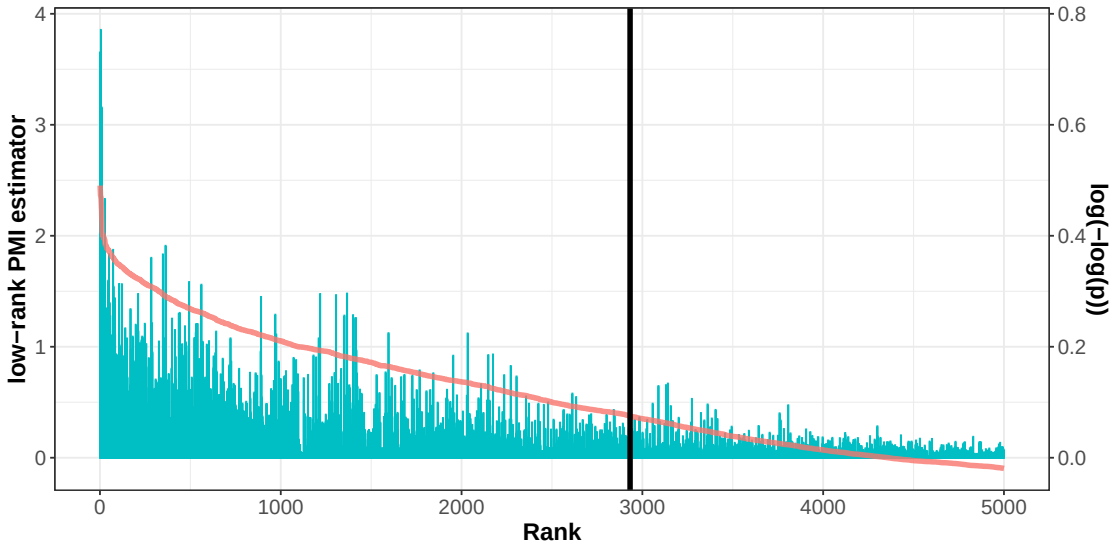

Figure 3: The estimated low-rank PMI of the top $5,000$ entities with the smallest p-values when quantifying their relationships with the Alzheimer's Disease (PheCode:290.11). The x-axis exhibits the rank of the p-values, the blue bars exhibit the estimated low-rank PMI and the red line exhibits $\log(-\log(p\text{-value}))$. The codes located to the left of the black line are identified as significantly related to the target feature, AD, at a target FDR of 0.01.

the smallest p-values in relation to AD. This visual representation highlights an important observation: entities w' with high $\widetilde{\mathbb{PMI}}_{w,w'}$ values do not necessarily correspond to small p-values. This finding underscores the limitations of existing methods that rely solely on the ranking of inner product values from KG embeddings to predict connections or edges between entities.

To validate the accuracy of the KG, we conduct a manual review of the top twenty entities with the smallest p-values. This group includes eight PheCodes (diseases) and twelve RXNORMs (drugs). These entities are enumerated in Table S2 in Supplementary S3. A substantial body of literature (Sosa-Ortiz et al., 2012; Bathgate et al., 2001; Nussbaum and Ellis, 2003; Reisberg et al., 2003; Tsuno, 2009; Olin and Schneider, 2002; Xu et al., 2008; Rabinowitz et al., 2007) confirms that these entities have a strong association with AD,

thereby validating the precision of our KG.

To further assess the effectiveness of the proposed algorithm in identifying entities associated with the given disease AD, we incorporate a comparison with advanced language models ChatGPT (OpenAI, 2023a) and GPT-4 (OpenAI, 2023b). Specifically, we first select the top 50 entities related to AD based on different methods, and then randomly sample an equal number of entities from the remaining set as negative samples. We then prompt ChatGPT and GPT-4 to rate the clinical relatedness between these 100 entities and the target disease AD, outputting a score on a scale of 0 to 1 for each entity. The evaluation involves computing the Spearman rank correlation between the negative p-values generated by our algorithm and the scores from ChatGPT/GPT4. Additionally, we examine correlations between cosine similarities obtained from alternative benchmarks and the ChatGPT/GPT scores, as detailed in Table 5. The results indicate that the rank correlation coefficients for KNIT with ChatGPT and GPT-4 are 0.553 and 0.582, respectively, outperforming other methods which all showed correlations less than 0.45 with the language models. For reference, the rank correlation between ChatGPT and GPT-4 themselves is 0.610.

	Statistics	KNIT	$\widehat{\mathbb{PMI}}$	Sap	PubMed	Bio	Bert
ChatGPT	cor	**0.553**	0.428	0.368	0.023	-0.267	-0.157
	p-value	0.001	0.001	0.001	0.397	0.996	0.945
GPT-4	cor	**0.582**	0.303	0.307	0.085	0.026	-0.152
	p-value	0.001	0.002	0.002	0.206	0.419	0.930

Table 5: Spearman rank correlation test between the score given by different methods and that given by ChatGPT for 100 pairs between AD and other clinical entities pairs. The correlation between the ChatGPT and GPT-4 is 0.610.

6 Conclusion

In this paper, we prove the asymptotic normality of the low-rank estimator $\widetilde{\mathbb{PMI}}$ and propose an algorithm for hypothesis testing on the low-rank KG $\mathbf{V}\mathbf{V}^\top$. When the patient-level data is approachable, the mean and variance of the entries of the knowledge matrix can be well estimated; when the patient-level data is not approachable, we estimate the covariance matrices for the low-rank estimator under the global null distribution, using them as a proxy for the actual covariance structure. The main idea of proposing the low-rank estimator is to utilize the low-rank structure of the KG and reduce the variance of the estimator to give a precise estimation of the KG.

References

Abdelaziz, I., A. Fokoue, O. Hassanzadeh, P. Zhang, and M. Sadoghi (2017). Large-scale structural and textual similarity-based mining of knowledge graph to predict drug–drug interactions. *Journal of Web Semantics 44*, 104–117.

Abu-Salih, B., M. AL-Qurishi, M. Alweshah, M. AL-Smadi, R. Alfayez, and H. Saadeh (2023). Healthcare knowledge graph construction: A systematic review of the state-of-the-art, open issues, and opportunities. *Journal of Big Data 10*(1), 81.

Ahuja, Y., L. Liang, D. Zhou, S. Huang, and T. Cai (2022, 02). Semisupervised Calibration of Risk with Noisy Event Times (SCORNET) using electronic health record data. *Biostatistics*.

Arora, S., Y. Li, Y. Liang, T. Ma, and A. Risteski (2016). A latent variable model ap-

proach to pmi-based word embeddings. *Transactions of the Association for Computational Linguistics 4*, 385–399.

Arora, S., Y. Li, Y. Liang, T. Ma, and A. Risteski (2018). Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association for Computational Linguistics 6*, 483–495.

Bathgate, D., J. Snowden, A. Varma, A. Blackshaw, and D. Neary (2001). Behaviour in frontotemporal dementia, Alzheimer's disease and vascular dementia. *Acta neurológica scandinavica 103*(6), 367–378.

Bauer-Mehren, A., P. Lependu, S. V. Iyer, R. Harpaz, N. J. Leeper, and N. H. Shah (2013). Network analysis of unstructured EHR data for clinical research. *AMIA Summits on Translational Science Proceedings 2013*, 14–18.

Beam, A. L., B. Kompa, A. Schmaltz, I. Fried, G. Weber, N. Palmer, X. Shi, T. Cai, and I. S. Kohane (2019). Clinical concept embeddings learned from massive sources of multimodal medical data. In *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2020*, pp. 295–306. World Scientific.

Benjamini, Y. and D. Yekutieli (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, 1165–1188.

Bordes, A., X. Glorot, J. Weston, and Y. Bengio (2014). A semantic matching energy function for learning with multi-relational data: Application to word-sense disambiguation. *Machine Learning 94*, 233–259.

Bordes, A., N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko (2013). Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems 26*.

Candès, E. J. and T. Tao (2010). The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory 56*(5), 2053–2080.

Carpentier, A., J. Eisert, D. Gross, and R. Nickl (2019). Uncertainty quantification for matrix compressed sensing and quantum tomography problems. In *High Dimensional Probability VIII: The Oaxaca Volume*, pp. 385–430. Springer.

Chen, I. Y., M. Agrawal, S. Horng, and D. Sontag (2019). Robustly extracting medical knowledge from EHRs: a case study of learning a health knowledge graph. In *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2020*, pp. 19–30. World Scientific.

Chen, Y., J. Fan, C. Ma, and Y. Yan (2019). Inference and uncertainty quantification for noisy matrix completion. *Proceedings of the National Academy of Sciences 116*(46), 22931–22937.

Chernozhukov, V., C. Hansen, Y. Liao, and Y. Zhu (2023). Inference for low-rank models. *The Annals of statistics 51*(3), 1309–1330.

Dai, Y., S. Wang, N. N. Xiong, and W. Guo (2020). A survey on knowledge graph embedding: Approaches, applications and benchmarks. *Electronics 9*(5), 750.

Davis, C. and W. M. Kahan (1970). The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis 7*(1), 1–46.

Devlin, J., M. Chang, K. Lee, and K. Toutanova (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*, pp. 4171–4186.

Finlayson, S. G., P. LePendu, and N. H. Shah (2014). Building the graph of medicine from millions of clinical narratives. *Scientific Data 1*, 140032.

Foucart, S., D. Needell, Y. Plan, and M. Wootters (2017). De-biasing low-rank projection for matrix completion. In *Wavelets and sparsity XVII*, Volume 10394, pp. 269–281. SPIE.

Gu, Y., R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare 3*, 1–23.

Harnoune, A., M. Rhanoui, M. Mikram, S. Yousfi, Z. Elkaimbillah, and B. El Asri (2021). Bert based clinical knowledge extraction for biomedical knowledge graph construction and analysis. *Computer Methods and Programs in Biomedicine Update 1*, 100042.

Hong, C., E. Rush, M. Liu, D. Zhou, J. Sun, A. Sonabend, V. M. Castro, P. Schubert, V. A. Panickan, T. Cai, et al. (2021). Clinical knowledge extraction via sparse embedding regression (KESER) with multi-center large scale electronic health record data. *NPJ digital medicine 4*(1), 1–11.

Jiang, J., T. Wang, B. Wang, L. Ma, and Y. Guan (2022). Gated tree-based graph attention network (gtgat) for medical knowledge graph reasoning. *Artificial Intelligence in Medicine 130*, 102329.

Lee, J., W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics 36*(4), 1234–1240.

Liu, F., E. Shareghi, Z. Meng, M. Basaldella, and N. Collier (2021). Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4228–4238.

Lu, J., J. Yin, and T. Cai (2023). Knowledge graph embedding with electronic health records data via latent graphical block model.

McInnes, B. T., T. Pedersen, and J. Carlis (2007). Using UMLS concept unique identifiers (CUIs) for word sense disambiguation in the biomedical domain. In *AMIA Annual Symposium Proceedings*, Volume 2007, pp. 533–537. American Medical Informatics Association.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems 26*, 3111–3119.

Nelson, C. A., R. Bove, A. J. Butte, and S. E. Baranzini (2022). Embedding electronic health records onto a knowledge network recognizes prodromal features of multiple sclerosis and predicts diagnosis. *Journal of the American Medical Informatics Association 29*(3), 424–434.

Nickel, M., V. Tresp, H.-P. Kriegel, et al. (2011). A three-way model for collective learning on multi-relational data. In *Icml*, Volume 11, pp. 3104482–3104584.

Nussbaum, R. L. and C. E. Ellis (2003). Alzheimer's disease and parkinson's disease. *New england journal of medicine 348*(14), 1356–1364.

Olin, J. and L. Schneider (2002). Galantamine for Alzheimer's disease. *The Cochrane database of systematic reviews* (3), CD001747–CD001747.

OpenAI (2023a). Chatgpt: optimizing language models for dialogue. https://openai.com/blog/chatgpt/.

OpenAI (2023b). Gpt-4 technical report. *ArXiv*. https://openai.com/research/gpt-4.

Rabinowitz, J., I. Katz, P. P. De Deyn, A. Greenspan, and H. Brodaty (2007). Treating behavioral and psychological symptoms in patients with psychosis of Alzheimer's disease using risperidone. *International psychogeriatrics 19*(2), 227–240.

Reisberg, B., R. Doody, A. Stöffler, F. Schmitt, S. Ferris, and H. J. Möbius (2003). Memantine in moderate-to-severe Alzheimer's disease. *New England Journal of Medicine 348*(14), 1333–1341.

Rotmensch, M., Y. Halpern, A. Tlimat, S. Horng, and D. Sontag (2017). Learning a health knowledge graph from electronic medical records. *Scientific reports 7*(1), 1–11.

Roy, A. and S. Pan (2021). Incorporating medical knowledge in bert for clinical relation extraction. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 5357–5366.

Sang, S., Z. Yang, L. Wang, X. Liu, H. Lin, and J. Wang (2018). Sematyp: a knowledge graph based literature mining method for drug discovery. *BMC bioinformatics 19*, 1–11.

Shang, Y., Y. Tian, M. Zhou, T. Zhou, K. Lyu, Z. Wang, R. Xin, T. Liang, S. Zhu, and J. Li (2021). EHR-oriented knowledge graph system: toward efficient utilization of non-used information buried in routine clinical practice. *IEEE Journal of Biomedical and Health Informatics 25*(7), 2463–2475.

Socher, R., D. Chen, C. D. Manning, and A. Ng (2013). Reasoning with neural tensor networks for knowledge base completion. *Advances in neural information processing systems 26*.

Sosa-Ortiz, A. L., I. Acosta-Castillo, and M. J. Prince (2012). Epidemiology of dementias and Alzheimer's disease. *Archives of medical research 43*(8), 600–608.

Tsuno, N. (2009). Donepezil in the treatment of patients with Alzheimer's disease. *Expert review of neurotherapeutics 9*(5), 591–598.

Wang, Z., J. Zhang, J. Feng, and Z. Chen (2014). Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the AAAI conference on artificial intelligence*, Volume 28.

Wen, J., X. Zhang, E. Rush, V. A. Panickan, X. Li, T. Cai, D. Zhou, Y.-L. Ho, L. Costa, E. Begoli, C. Hong, J. M. Gaziano, K. Cho, J. Lu, K. P. Liao, M. Zitnik, and T. Cai (2023, 02). Multimodal representation learning for predicting molecule–disease relations. *Bioinformatics 39*(2), btad085.

Xia, D. (2021). Normal approximation and confidence region of singular subspaces. *Electronic Journal of Statistics 15*(2), 3798–3851.

Xia, D., A. R. Zhang, and Y. Zhou (2022). Inference for low-rank tensors—no need to debias. *The Annals of Statistics 50*(2), 1220–1245.

Xu, H., H. Wang, L. Zhuang, B. Yan, Y. Yu, Z. Wei, Y. Zhang, L. E. Dyck, S. J. Richardson, J. He, X. Li, J. Kong, and X.-M. Li (2008). Demonstration of an anti-oxidative stress mechanism of quetiapine. *The FEBS Journal 275*(14), 3718–3728.

Yang, B., W.-t. Yih, X. He, J. Gao, and L. Deng (2014). Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575*.

Zhang, Y., M. Sheng, R. Zhou, Y. Wang, G. Han, H. Zhang, C. Xing, and J. Dong (2020). HKGB: an inclusive, extensible, intelligent, semi-auto-constructed knowledge graph framework for healthcare with clinicians' expertise incorporated. *Information Processing & Management 57*(6), 102324.

Zhou, D., T. Cai, and J. Lu (2023). Multi-source learning via completion of block-wise overlapping noisy matrices. *Journal of Machine Learning Research 24*(221), 1–43.

Zhou, D., Z. Gan, X. Shi, A. Patwari, E. Rush, C.-L. Bonzel, V. A. Panickan, C. Hong, Y.-L. Ho, T. Cai, et al. (2022). Multiview incomplete knowledge graph integration with application to cross-institutional EHR data harmonization. *Journal of Biomedical Informatics 133*, 104147.