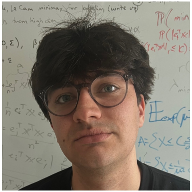
PsycheMERGE Platform Team: Knowledge Graph Tutorial

Parker Knight and Chenyin Gao

Harvard T.H. Chan School of Public Health

Our Team




Parker Knight
Harvard

Zhuoran Wei
Harvard

Chenyin Gao
Harvard

Rui Duan
Harvard

Acknowledgment: We thank Tianxi Cai's team for sharing resources with us.

Roadmap

Background

Knowledge graphs from EHR data

- Feature selection
- Data harmonization

Example

- Steps for building KG
- Steps for data harmonization

Summary

What is a knowledge graph?

A medical knowledge graph is a structured model that organizes various medical data as nodes (like symptoms, diseases, treatments) and relationships (edges) between them, integrating diverse health information into an interconnected network.

What is a knowledge graph?

A medical knowledge graph is a structured model that organizes various medical data as nodes (like symptoms, diseases, treatments) and relationships (edges) between them, integrating diverse health information into an interconnected network.

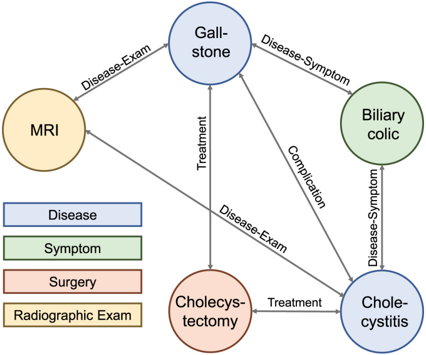


Figure: A small knowledge graph from Yang et al 2024

Knowledge graphs (cont.)

We can build a medical knowledge graph from all kinds of data sources.

Knowledge graphs (cont.)

We can build a medical knowledge graph from all kinds of data sources.

1. Medical literature: Zhang and Che 2021 construct a Parkinson's related KG from extensive literature review.

Knowledge graphs (cont.)

We can build a medical knowledge graph from all kinds of data sources.

1. Medical literature: Zhang and Che 2021 construct a Parkinson's related KG from extensive literature review.
2. Existing databases/LLM's: Google's Med-PaLM project.

Knowledge graphs (cont.)

We can build a medical knowledge graph from all kinds of data sources.

1. Medical literature: Zhang and Che 2021 construct a Parkinson's related KG from extensive literature review.
2. Existing databases/LLM's: Google's Med-PaLM project.

nature

Explore content ⌄ **About the journal** ⌄ **Publish with us** ⌄

nature > articles > article

Article | Open access | Published: 12 July 2023

Large language models encode clinical knowledge

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, ... Vivek Natarajan + Show authors

Nature **620**, 172–180 (2023) | Cite this article

285k Accesses | **1214** Altmetric | Metrics

Knowledge graphs (cont.)

We can build a medical knowledge graph from all kinds of data sources.

1. Medical literature: Zhang and Che 2021 construct a Parkinson's related KG from extensive literature review.
2. Existing databases/LLM's: Google's Med-PaLM project.
3. **Electronic health records.**

Roadmap

Knowledge Graph from EHR Data

Our objective is to develop a knowledge graph based on EHR data, where nodes represent EHR-derived terms, such as disease codes. Edges are established based on the likelihood of co-occurrence of these codes in the EHR data.

[1] Further details will be provided later.

Knowledge Graph from EHR Data

Our objective is to develop a knowledge graph based on EHR data, where nodes represent EHR-derived terms, such as disease codes. Edges are established based on the likelihood of co-occurrence of these codes in the EHR data.

We will define relationships between codes if they frequently co-occur within a specific time window, such as a month[1].

[1] Further details will be provided later.

EHR-derived Knowledge graph: an example

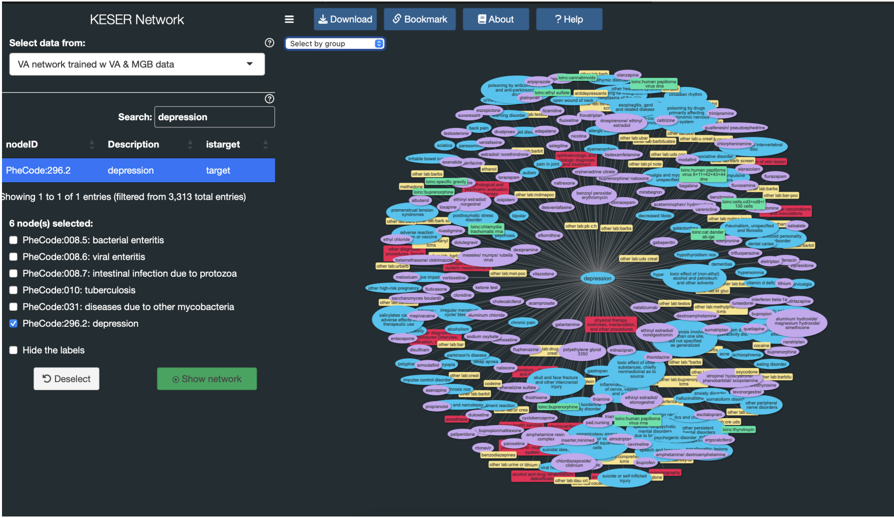


Figure: EHR-derived knowledge graph

EHR-derived Knowledge graph: use cases

EHR-based knowledge graphs summarize institutional medical knowledge. But how is this useful for our research at PsycheMERGE?

EHR-derived Knowledge graph: use cases

EHR-based knowledge graphs summarize institutional medical knowledge. But how is this useful for our research at PsycheMERGE?

1. Feature selection.
   - phenotyping model
   - risk prediction model
   - cohort definition
2. Cross-institutional data harmonization.
   - understand data heterogeneity, especially differences in coding behaviors
   - mapping code between institutions
   - learning common feature representations

Roadmap

Feature selection with EHR-KG

Say you want to build a prediction model for a particular disease using EHR code counts as predictors.

You may have count data for thousands of codes, many totally unrelated to the disease of interest.

A knowledge graph can be used to select codes that co-occur with the target disease, which can simplify the model building process.

Feature selection with EHR-KG

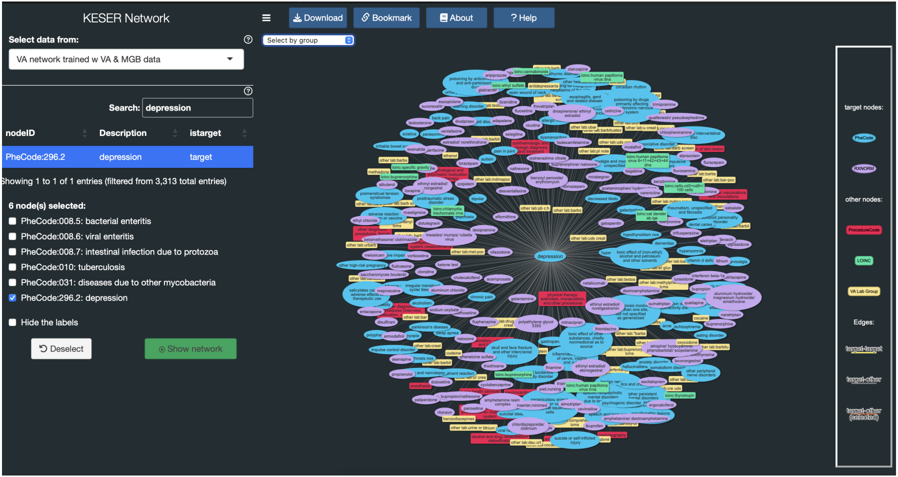
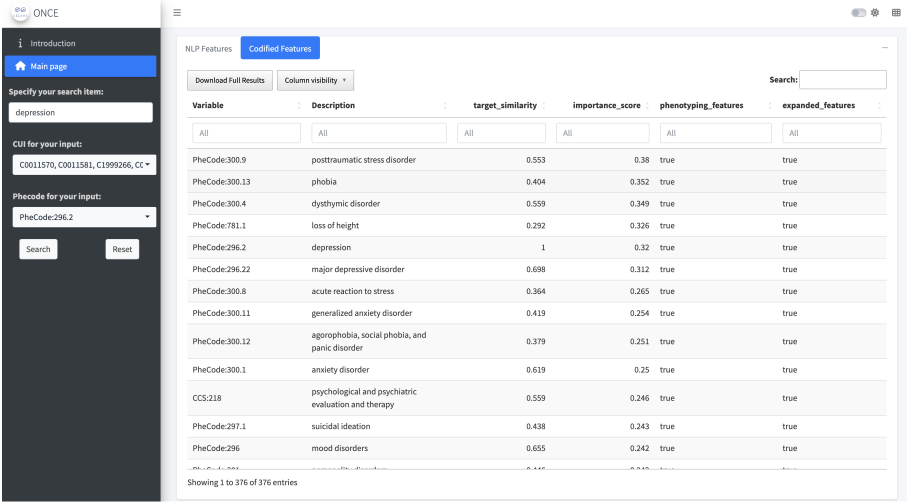


Figure: EHR-derived knowledge graph (depression as target term)

Feature selection with EHR-KG



Variable	Description	target_similarity	importance_score	phenotyping_features	expanded_features
PheCode:300.9	posttraumatic stress disorder	0.553	0.38	true	true
PheCode:300.13	phobia	0.404	0.352	true	true
PheCode:300.4	dysthymic disorder	0.559	0.349	true	true
PheCode:781.1	loss of height	0.292	0.326	true	true
PheCode:296.2	depression	1	0.32	true	true
PheCode:296.22	major depressive disorder	0.698	0.312	true	true
PheCode:300.8	acute reaction to stress	0.364	0.265	true	true
PheCode:300.11	generalized anxiety disorder	0.419	0.254	true	true
PheCode:300.12	agorophobia, social phobia, and panic disorder	0.379	0.251	true	true
PheCode:300.1	anxiety disorder	0.619	0.25	true	true
CCS:218	psychological and psychiatric evaluation and therapy	0.559	0.246	true	true
PheCode:297.1	suicidal ideation	0.438	0.243	true	true
PheCode:296	mood disorders	0.655	0.242	true	true

Figure: Features selected from the knowledge graph ranked by their relatedness to the target term (depression)

Roadmap

Background

Knowledge graphs from EHR data

- Feature selection
- Data harmonization

Example

- Steps for building KG
- Steps for data harmonization

Summary

Data harmonization with EHR-KG

Due to heterogeneity across healthcare systems, different providers may use different codes to represent the same underlying medical event (such as a diagnosis or procedure).

To responsibly leverage multi-site data, we need to correct for this by matching codes across sites.

Knowledge graphs summarize the co-occurrence information we need to do this.

Data harmonization: example

Using co-occurrence matrices from Site I and Site II, we can map codes from Site I to corresponding codes in Site II. This is achieved by aligning their embedding spaces and comparing the codes within this aligned space.

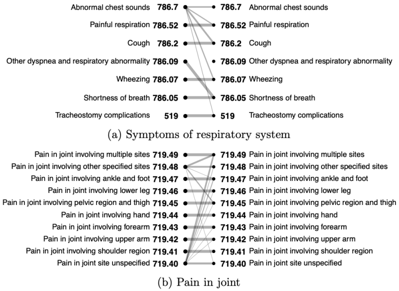


Figure 4: Plot of the estimated mapping of codes from VHA (left) to PHS (right). Selected codes belong to the group describing (a) symptoms of respiratory system, and (b) pain in joint. Line width indicates the magnitude of weight vector components.

Figure: Output of code matching algorithm from (Shi 2020).

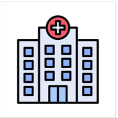
Roadmap

Steps for Building KG

- **Step 1**: Compute the site-specific co-occurrence matrix of *medical concepts* (e.g., PheCodes, Concept Unique Identifiers)
  - The co-occurrence can be defined in a given time window

Site 1

	C1	C2	C3
C1	4	4	4
C2		6	6
C3			8

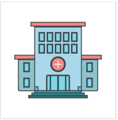


Site 2

	C1	C2	C3
C1	8	6	8
C2		6	6
C3			10

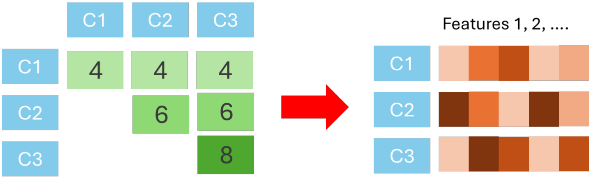
Site 3

	C1	C2	C3
C1	4	2	4
C2		2	2
C3			8



Steps for Building KG

- **Step 1**: Compute the site-specific co-occurrence matrix of *medical concepts* (e.g., PheCodes, Concept Unique Identifiers)
- **Step 2**: Compute the *concept embeddings*.
  - *Embeddings* are the summarized features

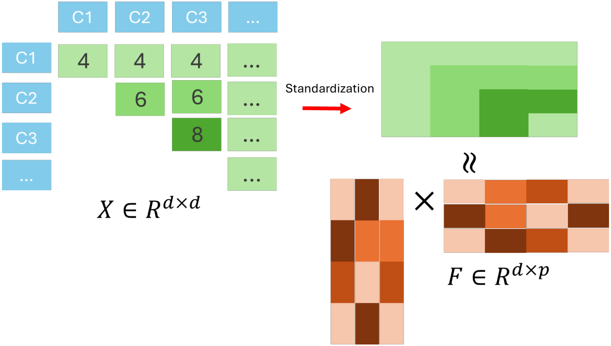


Site-specific Co-occurrence Matrix

Concept Embeddings

Steps for Building KG

- **Step 1**: Compute the site-specific co-occurrence matrix of *medical concepts* (e.g., PheCodes, Concept Unique Identifiers)
- **Step 2**: Compute the *concept embeddings*.
  - *Embeddings* are the summarized features
  - e.g., Singular Value Decomposition (SVD)



Steps for Building KG

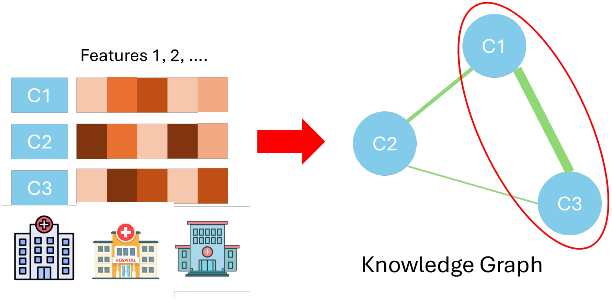
- **Step 1**: Compute the site-specific co-occurrence matrix of *medical concepts* (e.g., PheCodes, Concept Unique Identifiers)
- **Step 2**: Compute the *concept embeddings*.

Method	**Number of Concepts**
Code2Vec (Kartchner et al.; 2017)	8477
Med2Vec (Choi et al.; 2016)	28840
KESER (Hong et al.; 2021)	14718
MIKGI (Zhou et al.; 2022)	13261

Table: A List of Methods for Word Embeddings and Choosing their Dimensions

Steps for Building KG

- **Step 1**: Compute the site-specific co-occurrence matrix of *medical concepts* (e.g., PheCodes, Concept Unique Identifiers)
- **Step 2**: Compute the *concept embeddings*.
- **Step 3**: Compute the distances of the *concept embeddings*.
  - e.g., data-driven thresholding for feature selection



An example for co-occurrence matrix

- An example of visit-level data for codes C1 and C2

```
df_visit
#   Patient        Visit        Code
# 1       1   01/02/2025          C1
# 2       1   01/12/2025          C1
# 3       1   01/21/2025          C2
# 4       1   01/30/2025          C1
# 5       1   02/02/2025          C2
# 6       1   02/10/2025          C2
```

An example for co-occurrence matrix

- An example of visit-level data for codes C1 and C2

```
df_visit
#   Patient         Visit        Code
# 1       1    01/02/2025          C1
# 2       1    01/12/2025          C1
# 3       1    01/21/2025          C2
# 4       1    01/30/2025          C1
# 5       1    02/02/2025          C2
# 6       1    02/10/2025          C2
```

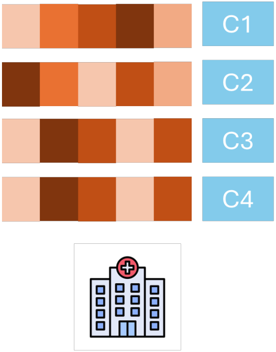
- Or simplified example (1 = "01/01/2025")

```
df_visit_simplified
#   Patient  Week        Code
# 1       1     2          C1
# 2       1    12          C1
# 3       1    21          C2
# 4       1    30          C1
# 5       1    33          C2
# 6       1    41          C2
```

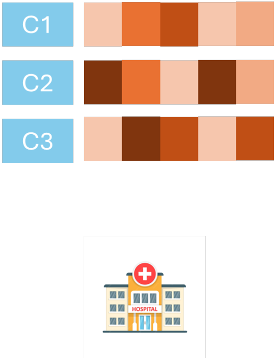
Roadmap

Steps for Data Harmonization

$$F_1 \in R^{d_1 \times p}$$

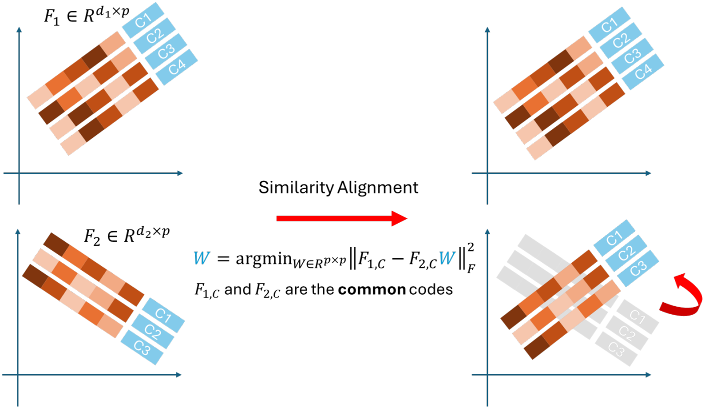


$$F_2 \in R^{d_2 \times p}$$



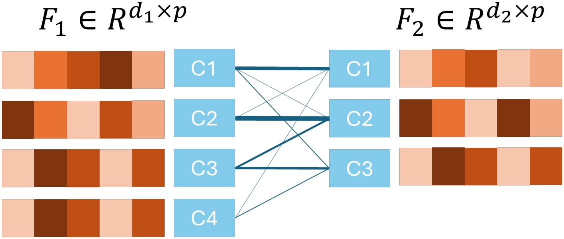
Steps for Data Harmonization

- **Step 1**: Space *alignment* for the *concept embeddings*
  - e.g., rotation-based space alignment



Steps for Data Harmonization

- **Step 1**: Space *alignment* for the *concept embeddings*
- **Step 2**: Code *mapping* within the **same group** (e.g., age strata, ancestry annotation).
  - e.g., top-K nearest neighbors matching



$$\Pi = \mathrm{argmin}_{\Pi \in R^{d_2 \times d_1}} \|F_1 - \Pi F_2 W\|_F^2$$

Roadmap

Background

Knowledge graphs from EHR data
- Feature selection
- Data harmonization

Example
- Steps for building KG
- Steps for data harmonization

Summary

Discussion

- **Effort from each site**
  - Data extraction & processing
  - Co-occurrence matrix calculation
- **Data sharing**
  - Co-occurrence matrix or embeddings
- **Cohort definition**
  - More inclusive (e.g., patients with at least one psychiatric visit)
- **Our group can offer**
  - Scripts & software for co-occurrence matrix
  - Troubleshooting support
  - Knowledge graph training & data harmonization upon receiving the matrix or embeddings
- **Timeline and team**

Summary

More info can be found below.

- Documentation on KG: https://cran.r-project.org/web/packages/kgraph/vignettes/kgraph.html
- Website App to build KG: https://celehs.connect.hms.harvard.edu/kesernetwork/

Thank you!

Appendix

An example (cont.)

- Summarize visit-level data into counts given a time window of **three**

```
df_visit
#   Patient Visit Parent_Code
# 1       1     1          C1
# 2       1     2          C1
# 3       1     3          C2
# 4       1     4          C1
# 5       1     5          C2
# 6       1     6          C2
window.length <- 3
df_count <- df_visit %>%
  mutate(Month = cut_interval(Visit, length = window.length)) %>%
  group_by(Patient, Parent_Code, Month) %>%
  summarise(Count = n())
df_count
#   Patient Parent_Code Month Count
# 1       1          C1 [0,3]     2
# 2       1          C1 (3,6]     1
# 3       1          C2 [0,3]     1
# 4       1          C2 (3,6]     2
```

An example (cont.)

```
df_count
#   Patient Parent_Code Month Count
# 1       1          C1 [0,3]     2
# 2       1          C1 (3,6]     1
# 3       1          C2 [0,3]     1
# 4       1          C2 (3,6]     2
```

- The co-occurred $\min(2, 1) = 1$ for patient 1 at the first **three** Visits

```
build_df_cooc(subset(df_count, Patient == 1 &
                     Month == '[0,3]'))
#>    C1 C2
#> C1  2  1
#> C2  .  1
```

An example (cont.)

```
df_count
#   Patient Parent_Code Month Count
# 1       1          C1 [0,3]     2
# 2       1          C1 (3,6]     1
# 3       1          C2 [0,3]     1
# 4       1          C2 (3,6]     2
```

- The co-occurred $\min(2, 1) = 1$ for patient 1 at the first **three** Visits

```
build_df_cooc(subset(df_count, Patient == 1 &
                     Month == '[0,3]'))
#>    C1 C2
#> C1  2  1
#> C2  .  1
```

- The co-occurred $\min(1, 2) = 1$ for patient 1 at the second **three** Visits

```
build_df_cooc(subset(df_count, Patient == 1 &
                     Month == '(3,6]'))
#>    C1 C2
#> C1  1  1
#> C2  .  2
```

An example (cont.)

- The co-occurrence matrix for patient 1 is

```
spm_cooc <- build_df_cooc(df_count)
spm_cooc
#>    C1 C2
#> C1  3  2
#> C2  .  3
```

- The overall visit-level co-occurrence is aggregated over all the patients

An example (cont.)

- The co-occurrence matrix for patient 1 is

```r
spm_cooc <- build_df_cooc(df_count)
spm_cooc
#>    C1 C2
#> C1  3  2
#> C2  .  3
```

- The overall visit-level co-occurrence is aggregated over all the patients
- The co-occurrence matrix is used
  1. to output the concept embeddings;
  2. to compute the correlation for the knowledge graph.

An example

- **Assumption**: codes that have more co-occurrence are more similar.

An example

- **Assumption**: codes that have more co-occurrence are more similar.
- Compute the PMI, a measure of similarity (such as SVD-SPPMI or GloVe)

```
spm_cooc
#>    C1 C2
#> C1  3  2
#> C2  .  3
m_pmi = get_pmi(spm_cooc)
m_pmi
#             C1         C2
# C1  0.1823216 -0.2231436
# C2 -0.2231436  0.1823216
```

where $\mathrm{PMI}(\text{Ci, Cj})$ is $\log\left\{\frac{P(Ci,Cj)}{P(Ci)P(Cj)}\right\}$ for $i, j = 1, 2$.

KG for EHR (cont.)

- Compute the SVD of the PMI matrix $\rightarrow$ embeddings

```
m_pmi
#            C1         C2
# C1  0.1823216 -0.2231436
# C2 -0.2231436  0.1823216
get_svd(m_pmi, 1)
#          [,1]
# C1  0.4502583
# C2 -0.4502583
```

where

1. the first k principal components are fed to the KG;
2. rank k can be chosen balancing the AUC and dimensionality.

Reference I

Choi, E., Bahadori, M. T., Searles, E., Coffey, C., Thompson, M., Bost, J., Tejedor-Sojo, J. and Sun, J. (2016). Multi-layer representation learning for medical concepts, *proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1495–1504.

Hong, C., Rush, E., Liu, M., Zhou, D., Sun, J., Sonabend, A., Castro, V. M., Schubert, P., Panickan, V. A., Cai, T. 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): 151.

Kartchner, D., Christensen, T., Humpherys, J. and Wade, S. (2017). Code2vec: Embedding and clustering medical diagnosis data, *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, IEEE, pp. 386–390.

Reference II

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