Appendix

1 Background

In this section, we review existing methods for knowledge graph representation learning [1], [6]. A KG is considered as a set of entities $\mathcal{E}$ and relations $\mathcal{R}$. The set of directed edges, $\mathscr{D}^+$ comprises triples (h, r, t) where a direction of relation r is from head h to tail t entity

TransE [1] is a simple and efficient translational based distance model. It models the relation as a translation vector between head and tail entity vectors. For the given two entity vectors $\boldsymbol{h}$, $\boldsymbol{t} \in \mathbb{R}^n$, it maps the relation as translation vector $\boldsymbol{r} \in \mathbb{R}^n$,i.e., $\boldsymbol{h} + \boldsymbol{r} \approx \boldsymbol{t}$ for observed triple $\boldsymbol{h}, \boldsymbol{r}, \boldsymbol{t}$. Thus, the distance based scoring function is defined as:

$$f(h, r, t) = \|\boldsymbol{h} + \boldsymbol{r} - \boldsymbol{t}\|_{l_1/l_2}, \tag{1}$$

where, $\|\cdot\|_{l_1/l_2}$ is the l_1 or l_2-norm of the difference vector. $f(h, r, t)$ will be minimized for plausible triples. A margin based pairwise ranking loss is used to differentiate between correct and incorrect triples by minimizing their TransE score difference. Formally, the loss function is defined:

$$\sum_{x \in \mathscr{D}^+} \sum_{y \in \mathscr{D}^-} \max(0, f(x) - f(y) + \gamma), \tag{2}$$

with respect to the entity and relation vectors. γ is a margin hyperparameter. $\mathscr{D}^+$ stores only positive triples, i.e., observed triples in KG. $\mathscr{D}^-$ is the set with negative examples that are drawn randomly.

Despite its simplicity and efficiency, TransE cannot model 1-to-N, N-to-1, and N-to-N type of relations as it does not learn a distributed representation of entities. To tackle these flaws, TransH [6] was introduced to model a relation r as a vector on a relation specific hyperplane, and project entities associated with it on the corresponding hyperplane for learning the entities' distributed representation.

2 Negative Sampling

For negative sampling, DARLING utilizes a uniform demographic agnostic approach, where it considers the set of all the triples that do not belong to the medical KG, irrespective of the demographic dimension. Formally, for the demographic set c, the negative samples are drawn from the set,

$$\begin{aligned} \mathscr{D}_c^- = &\{(h^{'}, r, t) | h^{'} \in \mathcal{E}, (h^{'}, r, t) \notin \mathscr{D}^+\} \\ &\cup \{(h, r, t^{'}) | t^{'} \in \mathcal{E}, (h, r, t^{'}) \notin \mathscr{D}^+\}. \end{aligned} \tag{3}$$

3 Medical Knowledge Graph Statistics

Our medical KG includes $9,289$ distinct entities and two types of relations – Disease_to_Treatment and Disease_to_Medicine. Regarding demographics, we end up with 79 different demographic set combinations (gender, age group and ethnic group). Finally, our KG contains $126,141$ distinct quadruples, $100,912$ of which we use for training, $10,091$ for validation and $15,138$ for testing. Table 1 gives details on the constructed KG.

Table 1. Medical KG number of entities, relations and demographics.

Entities		**Relations**		**Demographics**	
#Disease	6,968	#Disease_to_Treatment	58,225	#Gender	2
#Treatment	1,475	#Disease_to_Medicine	67,916	#Age group	6
#Medicine	846			#Ethnic group	7
#Total	9,289	#Total	126,141	#Total (sets)	79

4 Demographic Statistics

Table 2 illustrates the number of unique patients that belong to each demographic category. We constructed our medical KG using data from 46,520 patients and 58,976 admissions related to them. The grouping of age values (years) was done by us, considering that we wanted to distribute the patients equally in different groups. The genders and ethnic groups are adopted from MIMIC-III data [3].

Table 2. Demographic statistics for each category. Our medical KG contains data from 46,520 unique patients.

Gender	**#Patients**
male	26,121
female	20,399

Age Group	**#Patients**
[0-18)	7,942
[18-48)	7,005
[48-60)	7,515
[60-70)	7,860
[70-80)	7,939
>= 80	8,259

Ethnic Group	**#Patients**
white	32,372
black	3,871
asian	1,690
hispanic	1,642
native	46
other	1,489
unknown	5,410

Table 3. Detailed results of our experiments.

Task	Disease-Treatment				Disease-Medicine			
Methods	Mean Rank		Hits@10		Mean Rank		Hits@10	
TransE [1]	73.94		47.40%		27.04		54.33%	
TransH [6]	75.56		48.60%		27.71		55.46%	
TransR [5]	115.12		30.34%		45.74		39.16%	
TransD [2]	84.66		47.64%		33.51		55.76%	
PrTransE [4]	69.69		47.21%		27.51		54.80%	
PrTransH [4]	69.01		47.25%		26.71		55.73%	
Probability score	with	without	with	without	with	without	with	without
DARLING (Gender)	68.89	71.11	47.62%	45.83%	25.67	27.13	56.94%	54.58%
DARLING (Age)	66.46	69.32	50.48%	48.17%	23.84	25.16	59.71%	57.94%
DARLING (Ethnicity)	67.82	69.28	48.52%	46.85%	25.57	26.86	57.64%	55.42%
DARLING (G+A)	66.01	68.83	50.97%	48.17%	23.92	24.97	60.25%	58.09%
DARLING (G+E)	67.35	70.14	48.92%	45.96%	24.97	26.46	58.27%	56.12%
DARLING (A+E)	65.18	67.83	51.32%	48.25%	23.29	25.01	60.97%	59.31%
DARLING (all)	**64.65**	**67.18**	**52.19%**	**50.41%**	**22.86**	**24.89**	**61.73%**	**59.96%**

5 Model Configurations

For the experiments, we selected Adam optimizer, and we employ batch sizes of $b = \{128, 256, 512\}$, embedding dimensions of $d = \{128, 256, 512\}$, learning rates $lr = \{0.01, 0.001, 0.0001\}$, a margin $\gamma = 1$ and $p = 2$ for the scoring function. Furthermore, for the probabilistic hyper-parameters λ, e_p and e_n we use the values of 10^{-2}, 10^{-4}, and 10^{-15} respectively. We train DARLING for 100 epochs and select the best state by the corresponding lowest mean rank on the validation set.

6 Inference

For inference, we describe how DARLING can be used for medical recommendation tasks through a link prediction process. Given a query patient with demographic set $c \in \mathcal{C}$ (gender, age, ethnicity) and the query disease diagnosis $\mathscr{d} \in \mathcal{D}$, we use DARLING to project the disease $\mathscr{d}$ into the hyperplane w_c and recommend top-k treatments and medicines. More precisely, given a query $q = (c, \mathscr{d})$, for each treatment procedure $\forall p \in \mathcal{P}$ and medicine $\forall m \in \mathcal{M}$ we compute its triple score with $\mathscr{d}$ (i.e. $f_c(\mathscr{d}, r, p), f_c(\mathscr{d}, r, m)$) on the demographic hyperplane w_c, and then select the treatment p and medicine m with the top-k highest ranking scores as the recommendation.

7 Detailed Results

Table 3 presents detailed results of our framework. In particular, we illustrate results using all possible demographic category combinations, and we further provide results by including and excluding the probability scores. At the same time, we present the results of all other baselines. As we can see, DARLING outperforms all baselines when using all demographic categories and including the probability scores.

References

1. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. In: NIPS. Curran Associates, Inc. (2013)
2. Ji, G., He, S., Xu, L., Liu, K., Zhao, J.: Knowledge graph embedding via dynamic mapping matrix. In: 53rd ACL-IJCNLP. Association for Computational Linguistics (2015)
3. Johnson, A.E., Pollard, T.J., Shen, L., Li-Wei, H.L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., Mark, R.G.: Mimic-iii, a freely accessible critical care database. Scientific data (2016)
4. Li, L., Wang, P., Wang, Y., Wang, S., Yan, J., Jiang, J., Tang, B., Wang, C., Liu, Y.: A method to learn embedding of a probabilistic medical knowledge graph: Algorithm development. JMIR Med Inform (2020)
5. Lin, Y., Liu, Z., Sun, M., Liu, Y., Zhu, X.: Learning entity and relation embeddings for knowledge graph completion. In: 28th AAAI. AAAI Press (2015)
6. Wang, Z., Zhang, J., Feng, J., Chen, Z.: Knowledge graph embedding by translating on hyperplanes. In: 28th AAAI. AAAI Press (2014)