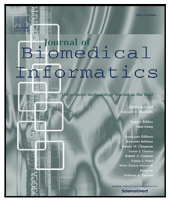
Journal of Biomedical Informatics 143 (2023) 104407

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin



Original Research

Graph-based clinical recommender: Predicting specialists procedure orders using graph representation learning

Sajjad Fouladvand [a,*,1], Federico Reyes Gomez [b,1], Hamed Nilforoshan [b], Matthew Schwede [a], Morteza Noshad [a], Olivia Jee [c], Jiaxuan You [b], Rok Sosic [b], Jure Leskovec [b], Jonathan Chen [a,d,e]

[a] Biomedical Informatics Research, Stanford University, Stanford, CA, United States of America
[b] Computer Science Department, Stanford University, Stanford, CA, United States of America
[c] Primary Care and Population Health, Stanford University, Stanford, CA, United States of America
[d] Division of Hospital Medicine, Stanford University, Stanford, CA, United States of America
[e] Clinical Excellence Research Center, Stanford University, Stanford, CA, United States of America

ARTICLE INFO

Keywords:
Graph neural networks
Electronic medical consultation
Hematology
Endocrinology

ABSTRACT

Objective: To determine whether graph neural network based models of electronic health records can predict specialty consultation care needs for endocrinology and hematology more accurately than the standard of care checklists and other conventional medical recommendation algorithms in the literature.

Methods: Demand for medical expertise far outstrips supply, with tens of millions in the US alone with deficient access to specialty care. Rather than potentially months long delays to initiate diagnostic workup and medical treatment with a specialist, referring primary care supported by an automated recommender algorithm could anticipate and directly initiate patient evaluation that would otherwise be needed at subsequent a specialist appointment. We propose a novel graph representation learning approach with a heterogeneous graph neural network to model structured electronic health records and formulate recommendation/prediction of subsequent specialist orders as a link prediction problem.

Results: Models are trained and assessed in two specialty care sites: endocrinology and hematology. Our experimental results show that our model achieves an 8% improvement in ROC-AUC for endocrinology (ROC-AUC = 0.88) and 5% improvement for hematology (ROC-AUC = 0.84) personalized procedure recommendations over prior medical recommender systems. These recommender algorithm approaches provide medical procedure recommendations for endocrinology referrals more effectively than manual clinical checklists (recommender: precision = 0.60, recall = 0.27, F1-score = 0.37) vs. (checklist: precision = 0.16, recall = 0.28, F1-score = 0.20), and similarly for hematology referrals (recommender: precision = 0.44, recall = 0.38, F1-score = 0.41) vs. (checklist: precision = 0.27, recall = 0.71, F1-score = 0.39).

Conclusion: Embedding graph neural network models into clinical care can improve digital specialty consultation systems and expand the access to medical experience of prior similar cases.

1. Introduction

Access to medical specialty care is often delayed due to growing limitations in clinicians' time and resources leading to over 25 million Americans with deficient access to specialty care [1], associated with higher mortality [2]. Prediction of medical procedures to be ordered during initial outpatient specialty consultation care can facilitate specialist consultations by eliminating delay and further follow-up of diagnostic steps, in some cases completely eliminating the need for an in-person consultation [3,4]. Clinical checklists are the current standard of practice to improve specialty referral healthcare delivery [5–8] and in critical situations such as those resulting from the COVID-19 pandemic [9].

Clinical checklists are labor-intensive to manually produce while both being not easily generalizable or personalizable to complex scenarios when they largely offer one-size-fits-all generic guidance and checklists for considerations [10–13]. Automated AI systems could

* Correspondence to: 3180 Porter Drive, Room A223.11, Palo Alto 94306, CA, United States of America.
@SajjadFV (S. Fouladvand).
E-mail addresses: sajjadf@stanford.edu (S. Fouladvand), frg100@alumni.stanford.edu (F.R. Gomez), hamedn@cs.stanford.edu (H. Nilforoshan), mschwede@stanford.edu (M. Schwede), noshad@stanford.edu (M. Noshad), ojee@stanford.edu (O. Jee), jiaxuan@cs.stanford.edu (J. You), rok@cs.stanford.edu (R. Sosic), jure@cs.stanford.edu (J. Leskovec), jonc101@stanford.edu (J. Chen).
[1] Authors contributed equally to this manuscript.

https://doi.org/10.1016/j.jbi.2023.104407
Received 23 March 2023; Received in revised form 15 May 2023; Accepted 25 May 2023
Available online 3 June 2023

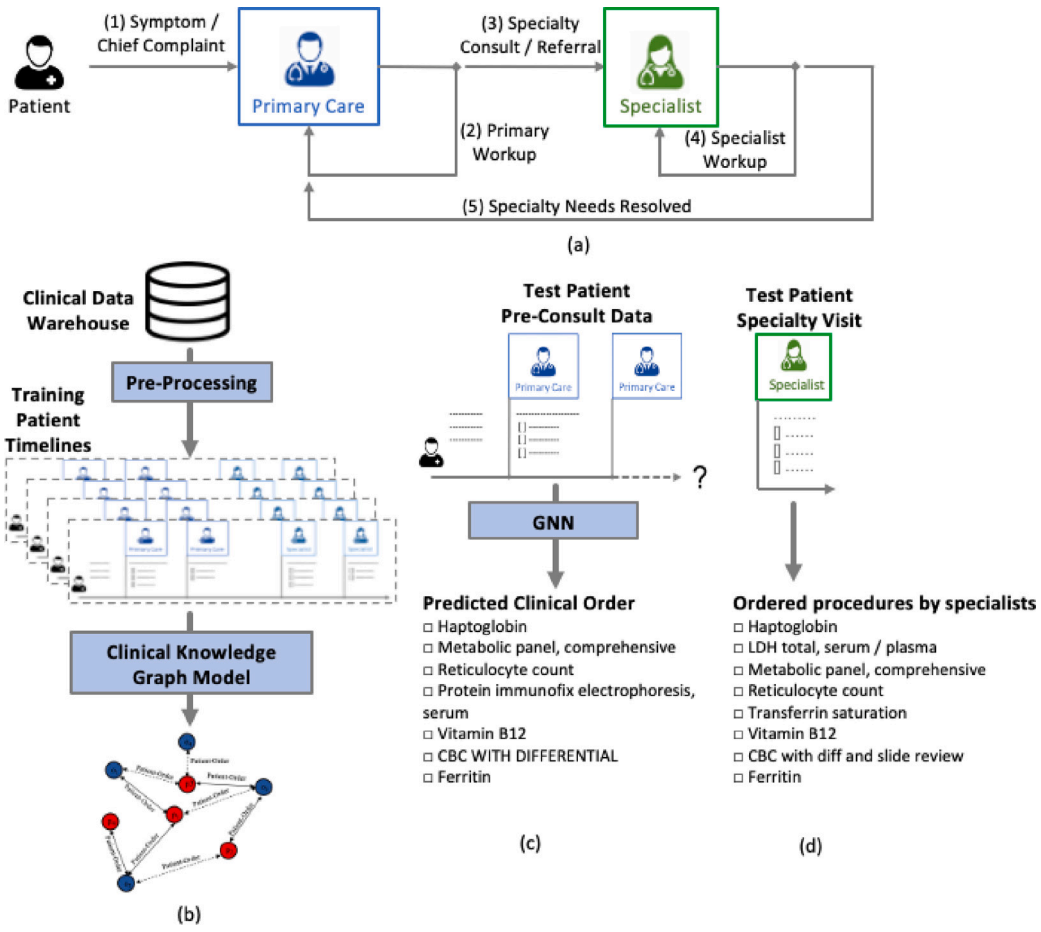


Fig. 1. Overall schema of the proposed framework. (a) Specialists' orders and workup can be initiated by primary care providers at referral or consultation time. (b) Patients' historical EHR data including diagnoses, procedures, and lab results were used to create a heterogeneous GNN. Nodes are patients and orders. Solid line edges show orders before referral and dashed line edges show the specialists orders after referral date. (c) Snapshot of the model recommendations for one patient. The GNN model was used to predict future specialists' orders. This figure shows the GNN model recommendations for one example patient referred to the hematology clinic. (d) Procedures ordered by specialists (ground truth) for the example patient shown in (c). This list highly overlaps with the procedures recommended by the GNN model (c).

improve specialty care systems by providing personalized recommendations based on prior subspecialist care, addressing limitations in general guidelines and checklists [14–16]. Leveraging artificial intelligence (AI) models trained using large scale routinely collected electronic health records (EHR) to create automatic specialty care procedure recommendation could improve efficient use of scarce clinician time and in turn increase access for more patients to reach appropriate care and consultation.

To this end, there have been multiple attempts to create automated and data-driven medical order recommender systems. OrderRex [17] was created based on association statistics and Bayesian rules to show promising results in improving clinical order decision making process and usability [10]. Ip et al. [18] used co-occurrence statistics to create a recommender algorithm to predict pediatric endocrinology patients' initial workup needs. Classical machine learning models [19], assessing coverage of manually authored order sets using optimization-based and clustering techniques [20], item-based collaborative filtering [21], and artificial neural networks were used for personalized general clinical orders [22] and endocrinology procedure recommendation [23].

To improve upon prior methods, we consider that the heterogeneity and structured nature of electronic health records (EHR) can be captured more effectively using graphical models [24–26]. A Graph Convolutional Transformer (GCT) [27] maps encounters into a fully connected graph and infers the underlying structure by computing self-attentions on the graph connection. Liu et al. [28] addressed the high visibility [29] of hub nodes such as demographic nodes and showed the effectiveness of modeling EHR data into heterogeneous graphs. Further, heterogeneous graph neural networks (GNN) have been utilized in drug pair side effect prediction [30], medical diagnosis prediction [31] and medical concept representations [32,33]. Graph structures are a widely adopted tool in the fields of medicine and biology. The ongoing progress in graph representation learning is expected to enhance the application of machine learning in healthcare [34]. Recently, there has been an increasing use of graph representation techniques in various medical domains, including computed tomography image analysis for COVID-19 [35] and pancreatic cancer diagnoses [36], hospital readmission prediction [37], drug discovery [38–40], and prescription recommendation using patients' prior encounters and lab tests [41]. The effectiveness of these approaches highlights the potential of graph representation methods in medical research and healthcare practice.

Motivated by Hamilton et al. [42], Zitnikv et al. [30], and Veličković et al. [43] we propose a novel GNN-based framework to provide personalized procedure order recommendations prior to or during patients' initial specialty care visits. Note, here we use the terminology 'order' to refer to the procedures ordered by physicians (e.g., laboratory tests, imaging studies, additional referrals and consultations).

We tested our models on Endocrinology and Hematology specialty referral care as two of the most common specialties that are receptive to virtual consultations given their strong basis in structured diagnostic test results [6,44,45]. Our objective is to determine whether GNN based models of EHR data can predict specialty consultation care

needs for endocrine and hematology more accurately than the standard of care guidelines and checklists and other conventional medical recommendation algorithms in the literature.

2. Materials and methods

Fig. 1 shows the overall schema of our proposed framework. We mapped patients' historical EHR data recorded prior to the patients' first referrals to specialty care clinics into a heterogeneous graph neural network. This model was trained to predict procedures ordered by endocrinology and hematology specialists during patients visits at the specialty care clinics.

2.1. Endocrinology data

Our data includes all outpatients referred by Stanford primary care providers to the Stanford Endocrinology clinic between January 2008 and December 2018. Use of this data for this study was approved by Stanford Institutional Review Board. We only included patients' first visit with the respective specialist within four months of their referral dates to reflect initial engagement with the specialist in response to the referral consultation request. Our final data set include 6821 new referrals to the endocrinology clinic.

We denote the list of patient referrals as $\boldsymbol{P} = \{p_1, \ldots, p_n\}$ in which n is the number of patient referrals. Each patient referral p_i constitutes a tuple $(t_i, \boldsymbol{D}^i, \boldsymbol{O}^i, \boldsymbol{L}^i, \boldsymbol{Y}^i)$, where t_i is referral's date and $\boldsymbol{D}^i \in \mathbb{R}^{10}$, $\boldsymbol{O}^i \in \mathbb{R}^{60}$, and $\boldsymbol{L}^i \in \mathbb{R}^{300\times3}$ are multi-hot encoded vectors representing diagnoses codes, procedure orders, and lab results for p_i recorded prior to t_i. We used a two month look back window for lab results and procedures. Each lab result was converted to a vector with three elements indicating (a) if p_i has had the lab result, (b) if the result was high, and (c) if the result was low. $\boldsymbol{Y}^i$ is a multi-hot encoded vector representing the procedures ordered by the specialist during patient's special care visit. Our final feature set includes 370 features. The target set includes 60 procedure orders. A full list of diagnoses, procedures and lab tests are presented in Table A1, Table A2, and Table A3 in Appendix A, respectively.

2.2. Hematology data

Our hematology cohort includes all outpatients referred to Stanford Hematology clinic by primary care providers as new patients from 2008 to 2021. Our final cohort includes 2007 patients. Data format is similar to the Endocrinology cohort described in Section 2.1. Feature set includes the top-100 most commonly recorded diagnoses, top-100 most commonly ordered labs as well as 33 medical procedures. The procedures were selected based on the health system's internally produced checklists for clinic referrals and electronic consultations. Lab tests were one-hot encoded based on their results and flagged as one of the following options: abnormal, normal, low, low off-scale, low panic, high, high off-scale, high panic, negative, and positive. Each Hematology patient referral p_i constitute $(t_i, \boldsymbol{D}^i, \boldsymbol{O}^i, \boldsymbol{L}^i, \boldsymbol{Y}^i)$, where t_i is referral's date and $\boldsymbol{D}^i \in \mathbb{R}^{100}$, $\boldsymbol{O}^i \in \mathbb{R}^{33}$, and $\boldsymbol{L}^i \in \mathbb{R}^{298\times3}$ are multi-hot encoded vectors representing diagnoses codes, procedure orders, and lab results for p_i prior to t_i. Table A4 in Appendix A lists the diagnoses, Table A5 shows the lab tests, and Table A6 shows the procedures used in this study for the Hematology cohort.

2.3. Descriptive analysis

Table 1 shows the cohort demographics. In general, demographics in terms of age, sex, race, and ethnicity were similar among endocrinology and hematology patients.

Tables 2 and 3 show the most common diagnoses, procedures and medications recorded for endocrinology and hematology referrals before patients' referral dates, respectively. Variable frequencies for each cohort were computed against new patients to all other specialty care clinics except our study cohorts. For our endocrinology cohort, the top-3 most frequently observed diagnosis codes were thyroid nodule, hyperthyroidism and hypothyroidism, and the top-3 most frequently ordered procedures were comprehensive metabolic panel, thyroid-stimulating hormone test (TSH), and T4 free. The Top-3 most frequently prescribed medications for this cohort were ondansetron, acetaminophen, and Normal Saline IV Bolus (reflecting supportive treatments that accompany specialty treatments like chemotherapy).

Anemia, thrombocytopenia and essential hypertension are the top-3 most frequently observed diagnosis codes recorded for hematology patients. Comprehensive metabolic panel, CBC, prothrombin time were most frequently ordered procedures and ondansetron, acetaminophen

Table 1
Patient characteristics among endocrinology and hematology cohorts. Numbers are N(25th percentile, 75th percentile) for age and N(%) for other variables.

Variable	Endocrinology (n=6,802)	Hematology (n=2007)
Age	52.85 (39, 66)	58.93 (46, 72)
Female	4,104 (66%)	1,066 (53%)
Race		
White	3,098 (50%)	1,131 (56%)
Other	1,083 (17%)	308 (15%)
Asian	1,420 (23%)	381 (19%)
Black	287 (5%)	122 (6%)
Unknown	245 (4%)	21 (2%)
Pacific Islander	82 (1%)	25 (1%)
Native American	23 (0.4%)	8 (0.4%)
Ethnicity		
Non Hisp./Lat.	5,251 (84%)	1,738 (86%)
Hisp./Lat.	737 (12%)	234 (12%)
Unknown	250 (4%)	35 (2%)

Table 2
Top-10 most observed diagnoses, procedures, and medications in endocrinology patients records. Count shows the number of encounters and Lift shows the ratio of frequency of each variable in the endocrinology cohort to its frequency in all specialty clinics except endocrinology.

Variable	Count	Lift
Diagnoses		
Thyroid nodule	422	17.01
Hyperthyroidism	403	60.87
Hypothyroidism	336	9.04
Diabetes mellitus	334	6.43
Essential hypertension	278	2.68
Osteoporosis	255	10.53
Hyperlipidemia	198	3.39
Neoplasm of thyroid	184	18.82
Vitamin D deficiency	149	4.04
Malaise and fatigue	131	4.63
Procedures		
Metabolic panel, comprehensive	1147	2.28
TSH	1121	4.27
T4, Free	744	8.06
Hemoglobin A1c	689	4.34
CBC with differential	492	1.71
Metabolic panel, basic	479	3.44
Vitamin D (25-Hydroxy)	365	2.73
ECG 12-lead	334	1.82
Lipid panel with calculated LDL	296	2.04
Magnesium	286	11.97
Medications		
Metformin	66	11.80
Diphth, pertus, tetanus	63	1.40
Pantoprazole	46	5.53
Metformin	43	10.35
Docusate sodium	41	14.66
Atorvastatin	39	6.82
Hydrocortisone	37	82.13
Levothyroxine	36	12.59
Sennosides	35	95.30
Insulin glargine	35	64.97

Table 3
Top-10 most observed diagnoses, procedures, and medications in hematology patients records. Count shows the number of encounters and Lift shows the ratio of frequency of each variable in the hematology cohort to its frequency in all specialty clinics except hematology.

Variable	Count	Lift
Diagnoses		
Anemia	365	37.66
Thrombocytopenia	345	45.92
Essential hypertension	200	2.7
Hyperlipidemia	147	3.51
Iron deficiency anemia	139	21.47
Pulmonary embolism and infarction	128	53.25
Leukopenia	118	75.44
Deep venous thrombosis of lower extremity	103	34.79
Diabetes mellitus	101	2.54
Shortness of breath	101	4.39
Procedures		
Metabolic panel, comprehensive	1,353	2.42
CBC with differential	937	2.6
Prothrombin time	825	8.79
ECG 12-lead	722	3.89
PTT partial thromboplastin time	614	13.31
Ferritin	580	8.2
Lactate Dehydrogenase (LDH)	567	14.89
Up ad lib	551	4143.86
Magnesium	549	22.46
Sequential compression device (SCD)	519	8.96
Medications		
Acyclovir	120	39.46
Pantoprazole	118	12.74
Polyethylene glycol	114	36.90
Sennosides	89	157.00
Docusate sodium	78	27.62
Acetaminophen	73	152.28
Oxycodone	72	12.33
Sulfamethoxazole-trimethoprim	70	7.95
Apixaban	61	46.12
Ferrous sulfate	58	37.08

and polyethylene glycol were the most frequently prescribed medications for hematology patients. Table 3 shows the full list of the top-10 most frequently observed diagnoses, procedures and medications in hematology patients.

2.4. Proposed method

2.4.1. Graph structure

We modeled patients' EHR data into a heterogeneous graph neural network $G = (V, E)$ (see Fig. 2(a)). V contains two node types: patient referral nodes $\{g^p_1, \ldots, g^p_{|\boldsymbol{P}|}\}$, and procedure order nodes $\{g^o_1, \ldots, g^o_{|\boldsymbol{O}|}\}$. Each patient node g^p_i is assigned a multi-hot encoded feature vector consisting of the concatenation of $\boldsymbol{D}^i$ and $\boldsymbol{L}^i$ and each procedure order node g^o_i are associated with one-hot encoding of the entity IDs (I_{s_i} and I_{o_i}, respectively).

Edge set E contains two edge types. 'ordered-with' edges with edge labels set to 0 that are edges between patient nodes and the procedures they have done before t_i, and 'ordered-with' edges with edge label set to 1 that connect the patients with the procedures that their specialist ordered during the specialty care visit after t_i. Note, 'ordered-with' edges with edge labels equal to 1 that represent specialist's orders after referral date were not used during training and were only used in the prediction phase as we are aiming to predict procedure orders after t_i. We formulate this task as binary link prediction of the existence of 'ordered-with' edges between a patient and an order. Further, node degree, node clustering coefficient and centrality transformations were applied to add synthetic features to each node feature vector. While the model can learn those features on its own, we added them to help the model focus on learning other features. We apply a different graph convolutional layers with independent parameters to each message type of (head, relation, tail) and aggregate embeddings across all node types. The same graph attention mechanism was applied to all node types.

2.4.2. Message passing and graph attention

Fig. 2(b) shows our proposed architecture. A fully connected layer with hidden size of 128 was used to map each node feature vector to pre-embedding vectors. Distinct fully connected layers were used for each node type. Two message passing layers were used each consisting of a dropout layer, a PReLU activation function, and a graph convolutional layer.

A custom heterogeneous graph attention layer was used using 1-head attention mostly following the structure of the original graph attention networks [43], with the following modifications: (1) we applied fully connected layers with batch normalization to the node embeddings and the neighbor embeddings, and (2) we aggregated neighbor embeddings using the attention mechanism and concatenated the aggregated embedding to the current node's embedding. This is then passed into a fully connected layer that reduces this down to a single output embedding followed by a batch normalization operation. Eq. (1) shows our message passing function.

$$\begin{aligned} x_v^{(1)} &= \mathrm{MLP}(x_v^{(0)}) \\ x_v^{(2)} &= \mathrm{GATConv}(\mathrm{PReLU}(\mathrm{Dropout}(x_v^{(1)}))) \\ x_v^{(3)} &= \mathrm{GATConv}(\mathrm{PReLU}(\mathrm{Dropout}(x_v^{(2)})) \\ &\quad + \beta * x_v^{(0)}) \\ x_v^{(4)} &= \mathrm{MLP}(x_v^{(3)}) \end{aligned} \tag{1}$$

and Eq. (2) shows the GATConv update function

$$\begin{aligned} \mathrm{aggr} &= \Sigma_{v_o \in \mathcal{N}(v)} \alpha_{v_o} * \mathrm{MLP}(x_{v_o}^{(k)}) \\ x_v^{(k+1)} &= \mathrm{MLP}(\mathrm{aggr} + \mathrm{MLP}(x_v^{(k)})) \end{aligned} \tag{2}$$

Where α_{v_o} is the 1 head GAT attention score calculated for v_o, $\mathcal{N}(v)$ is neighbors of v, and $x_v^{(0)}$ represents the node features of node v. The final predictions on existence of an 'ordered-with' edge e_{ij} between nodes g_i^p and g_j^o is inferred by concatenating their node embeddings and passing that through a fully connected two-layer perceptron, a batch normalization, a ReLU activation, and a final fully connected layer that outputs 2-dimensional logit vectors that are converted to final binary predictions using a softmax function. The formula for the link prediction head is as follows:

$$p = \mathrm{FC}(\mathrm{ReLU}(\mathrm{BN}(\mathrm{FC}([x^{(4)}_{g_i^p}; x^{(4)}_{g_j^o}])))) \in \mathbb{R}^2 \tag{3}$$

where BN refers to Batch Normalization and the first value corresponds to the probability that the edge exists and the second that it does not.

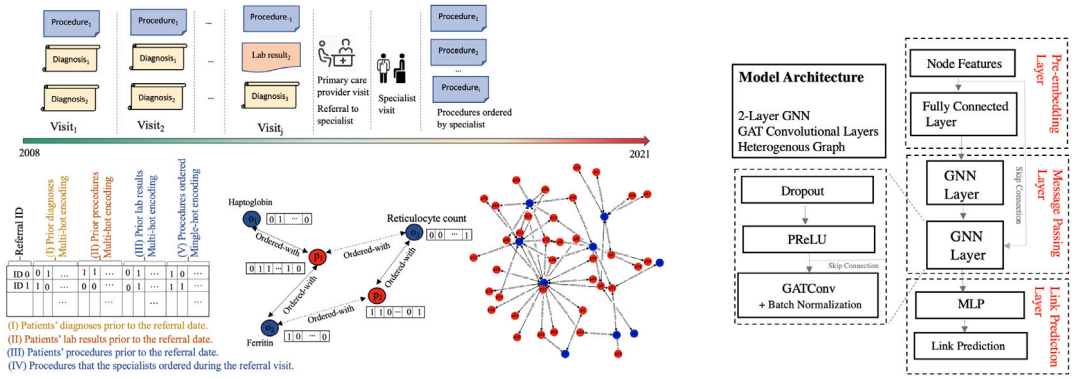
3. Experimental results

3.1. Endocrinology medical procedure recommendation

We used transductive disjoint training with a 1:4 positive:negative sampling. GNN models were implemented in Python and using Pytorch [46], Pythorch Geometric (PyG) [47,48] and DeepSNAP libraries [49,50].

Randomized cross validation with a set of 30 randomly selected hyper-parameter sets were used to tune the model. The final model uses Adam optimizer, dropout of 0.2, GAT convolutional layer, skip connection, and learning rate, hidden size, embedding size, and pre-embedding size are 1e-3, 64, 128 and 64, respectively. The model was tested on predictions made on all 'ordered-with' edges between a patient and an order placed during specialty visit and on an unseen test set consisting of 1, 321 patients.

Table 4 compares prediction results of our proposed GNN model with the baselines presented by Noshad et al. [23] including fully connected multi-layer neural network (Diagnostic Model), a collaborative filtering auto-encoder (AE), singular value decomposition (SVD),



(a) Graph construction. (b) Model architecture.

Fig. 2. Patients EHR data are formatted in a multi-hot encoding matrix. Red nodes show patients and blue nodes show procedures. Each patient node is assigned with a multi-hot encoded feature vector consisting of the concatenation of diagnoses and lab results features, and each procedure order node are associated with one-hot encoding of the procedure IDs. Solid edges show procedures ordered before referral dates and dashed edges show procedures ordered by specialists and after referral dates (targets). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
Performance of endocrinologist procedure order prediction models.

Model	AUC ROC	P@R 0.50	P@R 0.40	P@R 0.30	MSE	RMSE
Diagnostic Model	0.65	0.33	0.42	0.46	0.09	0.30
AE	0.73	0.23	0.33	0.49	0.07	0.26
PMF	0.62	0.22	0.31	0.43	0.06	0.25
SVD	0.74	0.23	0.33	0.50	0.06	0.24
Aggregated NN	0.73	0.31	0.41	0.53	0.08	0.28
Ensemble Model	0.80	0.37	0.47	0.57	0.06	0.25
GNN	0.88	0.42	0.49	0.57	0.05	0.23

probabilistic matrix factorization (PMF), an aggregate neural networks (Aggregated ANN), and an ensemble model (Ensemble Model) that uses a multi-layer neural network to combine the outputs of the diagnostic model, the collaborative filtering auto-encoder and the specialists' identifiers as a separate input signal.

Our proposed model can predict endocrinology specialty procedure orders for the new patient referrals more effectively (ROC-AUC = 0.88) compared to all models evaluated by Noshad et al. [23] (best ROC-AUC of the baselines = 0.80). Further, our model showed significantly higher precision at recalls 0.5, 0.4 and 0.3, and lower mean squared error (MSE) compared to all baseline models. Note, we used the same data as the data that were used in [23] except we removed features related to the specialists that patients were referred to, because although incorporating specialists' information in the model can lead to even higher accuracy, the information on specific specialists in the clinic can add bias to the model.

Further, we compared our proposed GNN model with clinical checklist for endocrinology procedure recommendation. The proposed recommender algorithm approach provides medical procedure recommendations for endocrinology referrals more effectively than manual endocrinology checklists (recommender: precision = 0.60, recall = 0.27, F1-score = 0.37) vs. (checklist: precision = 0.16, recall = 0.28, F1-score = 0.20).

3.2. Hematology medical procedure recommendation

We compared our proposed GNN model's performance in hematology procedure order recommendation with classical neural network and collaborative filtering based methods described in the previous section including a fully connected multi-layer neural network (Diagnostic Model), collaborative filtering auto-encoder (AE), singular value decomposition (SVD), probabilistic matrix factorization (PMF), aggregated neural networks (ANN), and an Ensemble Model combining the diagnostic model and the collaborative filtering auto-encoder. Model tuning follows a randomized cross validation method similar to the model trained for the endocrinology cohort. Our optimized GNN model has GAT convolutions and the hidden size, embedding size, pre-embedding size and learning rate are 32, 128, 32, 1e-3, respectively. All models are tested using an unseen test set including 603 patients. The comparison results are presented in Table 5. Our proposed GNN model predicts procedures ordered during patients first visit with hematology clinic at least 5% more effectively in terms of ROC-AUC (ROC-AUC = 0.84) compared to all baseline models. Further, the proposed model has higher precision at recalls 0.50 (precision = 0.41) and 0.40 (precision = 0.44) compared to the baselines.

Further, we compared our proposed GNN model with the clinical checklist in active use by the healthcare system's electronic consultation program. This checklist was produced by clinical committees in the health system to guide primary care providers when creating virtual consultations for Stanford hematology department. The checklist is offered for referral diagnoses including anemia, isolated erythrocytosis, elevated ferritin, isolated leukocytosis, isolated leukopenia, mgus, thrombocytopenia, thrombocytosis, and VTE/thrombophilia.

We compared the prediction performance of the proposed model with Stanford hematology electronic consultation checklist. The GNN recommender algorithm approaches provide medical procedure recommendations for hematology referrals more effectively than manual clinical checklists in terms of precision and F1-score (GNN recommender: precision = 0.44, recall = 0.38, F1-score = 0.41) vs. (checklist: precision = 0.27, recall = 0.71, F1-score = 0.39). We extracted referral diagnoses for the patients in our testing set and used the clinical checklist to predict the procedures they will need and compared these guideline suggestions against the procedures actually ordered during patients' first visits at the hematology clinic as the ground truth. Utilizing the clinical checklist is not possible for many patients because the guideline does not cover all referral diagnoses in the data. As a result, the test set for guideline is a subset of our original test set including 315 patients, and these results thus overestimate the performance of

Table 5
Performance of the models on medical procedure order recommendation for new referrals to hematology department.

Model	AUC ROC	P@R 0.50	P@R 0.40	P@R 0.30	MSE	RMSE
Diagnostic Model	0.60	0.18	0.19	0.21	0.21	0.45
AE	0.49	0.17	0.22	0.33	0.15	0.38
PMF	0.38	0.10	0.16	0.22	0.11	0.34
SVD	0.74	0.38	0.43	0.46	0.10	0.31
Aggregated NN	0.64	0.24	0.27	0.32	0.17	0.41
Ensemble Model	0.79	0.40	0.41	0.41	0.09	0.30
GNN	0.84	0.41	0.44	0.45	0.09	0.31

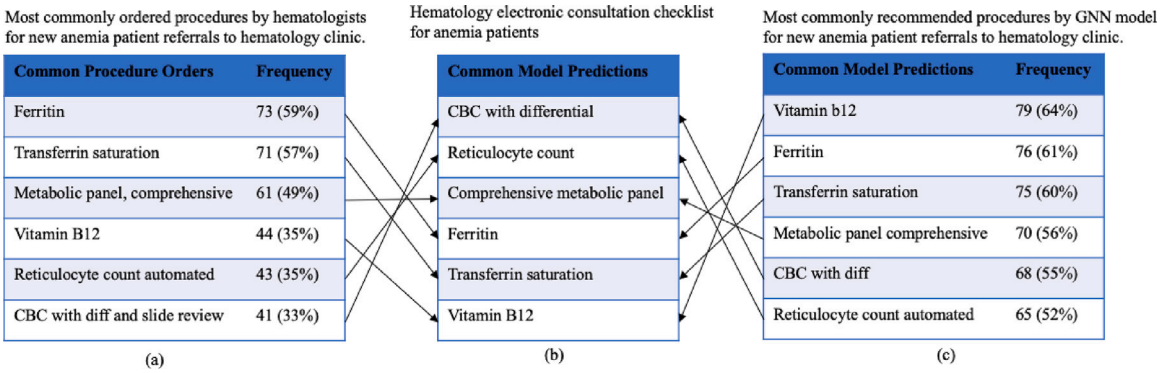


Fig. 3. (a) Procedures commonly ordered by hematologists for new anemia patient referrals to the hematology clinic. (b) Procedures recommended by the hematology checklist for anemia referrals (c) GNN most common recommendations for new anemia patient referrals to the hematology clinic. Arrows connect similar procedures. Model predictions, hematology checklist, and procedures ordered by hematologists overlap indicates the consistency of the proposed GNN model decision making with the existing clinical checklist and the ground truth.

the guideline checklists given that they would not perform at all in the cases without a clear matching referral diagnosis that our recommender algorithm models are always able to adapt to.

We further explored the GNN model and hematologists behaviors for an example subset of the patients in our testing set who were referred to the hematology clinic with an anemia diagnosis (124 patients) as the most common referral diagnosis in our data. Figure A1 shows the top six procedures that were most frequently ordered by hematologists which naturally overlaps with the six procedures provided in the hematology checklist. A list of six procedures that were commonly recommended by our GNN model is also shown in Fig. 3. Model's predictions overlap with both the hematology checklist suggestions (vitamin B12, ferritin, transferrin saturation, CMP, CBC with differential, and reticulocyte count) as well as the commonly ordered procedures by hematologists including ferritin, transferrin saturation, metabolic panel, comprehensive, vitamin B12, reticulocyte count automated, and CBC with diff and slide review. Although, given the high accuracy of the model this was expected, this provides some explanations on the proposed model's performance and shows that our GNN model's recommendations are consistent with the guideline as well as the ground truth (the procedures ordered by the specialists). We further explored the target variables (procedures) where the model had highest true positives. The top 5 procedures (targets) which had the highest true positives (model recommended them correctly) include comprehensive metabolic panel, reticulocyte count, ferritin, transferrin saturation, and vitamin B12. Excluding the most commonly predicted targets by the model, procedures such as CBC with differential, transferrin saturation, and haptoglobin had the highest false positive rates (model recommended but not actually ordered in subsequent specialist visits).

4. Discussion

In this study, we proposed a novel graph neural network based framework for medical procedure recommendation for specialty referral and virtual consultations. Models were trained and tested using new patients' referrals to endocrinology and hematology clinics as two of the most common specialties with frequent consultation requests. Patients' historical electronic health records were used to extract the predictors and the problem was models as a link prediction task in a heterogeneous graph structure. Proposed graph neural network based framework outperforms similar endocrinology and hematology medical procedure recommender algorithms in the literature in terms of ROC-AUC, precision, recall and f1-score. The GNN model outperforms endocrinology clinical checklists in terms of precision, recall and f1-score, and outperforms hematology clinical checklist in terms of precision and f1-score.

Clinical checklists are often limited to pre-defined set of diagnoses which precludes them from being used for a large group of patients with referral diagnoses outside of the scope of the guidelines. Our proposed AI models are end to end models that can analyze entire patient histories of EHR data and provide personalized recommendations. Using automated medical recommender tools could improve access to medical consultation guidance to patients by reducing the labor for clinicians and provide a consistent decision making support system for endocrinologists and hematologists and help them manage the ever-escalating complexity of electronic health records and medical guidelines. Such an automated medical recommender can provide faster and more accurate decision-making, improved identification of patterns and correlations in medical data, reduced human error, increased consistency in medical decision-making, continuous updating based on new medical data, and optimized medical resource allocation. The use of automated AI in medical procedure recommendation has the potential to improve healthcare quality, reduce medical errors, and increase efficiency.

Primary care providers could also benefit from an order recommender system that suggests orders that sub-specialists might place, anticipating patient needs without time delays and space separation that both reduce access for vulnerable patient population. This study showed the opportunities and a pathway towards such an automated medical procedure recommender system. Further, our proposed medical procedure recommender can assist primary care providers ordering electronic consultations to specialty care clinics. Our model can enrich

electronic consultations by providing detailed procedure recommendations leading to a more accurate and informative query from primary care providers and increasing the likelihood of specialists responding to consultations with timely and effective advice.

Limitations in the study include that the models were built as an outpatient recommender system, but many of the features (notably hematology related chemotherapy support orders) were based in the inpatient setting. This may have implications for generalizability in settings where in-patient records are not as easily accessible. Therefore, available clinical checklists such as order set templates remain valuable for specialty care settings. It is worth noting that the current models were trained to recommend procedures ordered by specialists. However, these targets may include noise and human errors and may not necessarily be correct and accurate. More follow-up studies needed for outpatient based outcomes and reinforcement learning towards crowdsourced recommender items and literature evidence-based clinical practice checklist as well as association with patients outcomes

AI performance in personalized recommendation for medical procedures for endocrinology and hematology patients shows the potential of combining both AI and manual approaches to help primary care providers when referring patients for specialty care or requesting virtual consultation.

5. Conclusion

In conclusion, our study proposed a novel framework based on a heterogeneous graph neural network for specialty care medical procedure recommendations for new patient referrals. Our proposed framework demonstrated superior performance compared to existing clinical checklists. The implementation of GNN models enabled accurate predictions, allowing for more effective management of patient referrals and enhanced overall quality of patient specialty care. Our findings suggest the potential of GNN models as a promising approach for medical procedure recommendation systems in endocrinology and hematology specialty care.

CRediT authorship contribution statement

Sajjad Fouladvand: Conceptualized the study, Designed the methodology produced the models, Collected and analyzed the data, Drafted the original manuscript. **Federico Reyes Gomez:** Conceptualized the study, Designed the methodology produced the models, Collected and analyzed the data, Drafted the original manuscript. **Hamed Nilforoshan:** Methodology and model development, Provided valuable input for reviewing and editing the manuscript. **Matthew Schwede:** Data collection, Clinical checklist development, Clinical validation of the models. **Morteza Noshad:** Methodology and model development, Provided valuable input for reviewing and editing the manuscript. **Olivia Jee:** Data collection, Clinical checklist development, Clinical validation of the models. **Jiaxuan You:** Methodology and model development, Provided valuable input for reviewing and editing the manuscript. **Rok Sosic:** Methodology and model development, Provided valuable input for reviewing and editing the manuscript. **Jure Leskovec:** Supervision throughout the project, Managed project administration, tasks, Offered valuable input during the reviewing and editing stages. **Jonathan Chen:** Supervision throughout the project, Managed project administration, tasks, Offered valuable input during the reviewing and editing stages.

Declaration of competing interest

Sajjad Fouladvand serves as a consultant for Roche. Jonathan Chen reported receiving consulting fees from Suton Pierce and Youneer Hyde MacFarlane PLLC and being a co-founder of Reaction Explorer LLC, a company that develops and licenses organic chemistry education software using rule-based artifcial intelligence technology.

Acknowledgments

All authors have reviewed and approved the final version of the manuscript and take responsibility for the accuracy and integrity of the work presented in this publication. We acknowledge the following fundings that supported this research: NIH/NLM (R56LM013365), Stanford Center for Artificial Intelligence in Medicine and Imaging, Doris Duke Charitable Foundation-Covid-19 Fund to Retain Clinical Scientists (20211260), Google Inc (SPO136094), and the American Heart Association - Strategically Focused Research Network.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jbi.2023.104407.

References

[1] Steffie Woolhandler, David U. Himmelstein, The relationship of health insurance and mortality: is lack of insurance deadly? Ann. Internal Med. 167 (6) (2017) 424–431.
[2] Julia C. Prentice, Steven D. Pizer, Delayed access to health care and mortality, Health Serv. Res. 42 (2) (2007) 644–662.
[3] Jonathan Chiang, Andre Kumar, David Morales, Divya Saini, Jason Hom, Lisa Shieh, Mark Musen, Mary K Goldstein, Jonathan H Chen, Physician usage and acceptance of a machine learning recommender system for simulated clinical order entry, AMIA Summits Transl. Sci. Proc. 2020 (2020) 89.
[4] Judy E Kim-Hwang, Alice Hm Chen, Douglas S Bell, David Guzman, Hal F Yee, Margot B Kushel, Evaluating electronic referrals for specialty care at a public hospital, J. Gen. Intern. Med. 25 (10) (2010) 1123–1128.
[5] Kathryn L Fantasia, Mary-Catherine Stockman, Zhihui Ju, Paola Ortega, Erika L Crable, Mari-Lynn Drainoni, Allan J Walkey, Megan Bergstrom, Katelyn O'Brien, Devin Steenkamp, Professional continuous glucose monitoring and endocrinology eConsult for adults with type 2 diabetes in primary care: results of a clinical pilot program, J. Clin. Transl. Endocrinol. 24 (2021) 100254.
[6] Erin Keely, Clare Liddy, Amir Afkham, Utilization, benefits, and impact of an e-consultation service across diverse specialties and primary care providers, Telemed. E-Health 19 (10) (2013) 733–738.
[7] Barbara J. Siepierski, Electronic consult experience: making health care more accessible and convenient for veterans, Fed. Pract. 30 (12) (2013) 38–40.
[8] Varsha G Vimalananda, Gouri Gupte, Siamak M Seraj, Jay Orlander, Dan Berlowitz, Benjamin G Fincke, Steven R Simon, Electronic consultations (e-consults) to improve access to specialty care: a systematic review and narrative synthesis, J. Telemed. Telecare 21 (6) (2015) 323–330.
[9] Paul Webster, Virtual health care in the era of COVID-19, Lancet 395 (10231) (2020) 1180–1181.
[10] Andre Kumar, Rachael C Aikens, Jason Hom, Lisa Shieh, Jonathan Chiang, David Morales, Divya Saini, Mark Musen, Michael Baiocchi, Russ Altman, et al., OrderRex clinical user testing: a randomized trial of recommender system decision support on simulated cases, J. Am. Med. Inform. Assoc. 27 (12) (2020) 1850–1859.
[11] Pierluigi Tricoci, Joseph M Allen, Judith M Kramer, Robert M Califf, Sidney C Smith, Scientific evidence underlying the ACC/AHA clinical practice guidelines, JAMA 301 (8) (2009) 831–841.
[12] Ron C Li, Jason K Wang, Christopher Sharp, Jonathan H Chen, When order sets do not align with clinician workflow: assessing practice patterns in the electronic health record, BMJ Qual. Safety 28 (12) (2019) 987–996.
[13] Elizabeth A McGlynn, Steven M Asch, John Adams, Joan Keesey, Jennifer Hicks, Alison DeCristofaro, Eve A Kerr, The quality of health care delivered to adults in the United States, N. Engl. J. Med. 348 (26) (2003) 2635–2645.
[14] B. Middleton, D.F. Sittig, A. Wright, Clinical decision support: a 25 year retrospective and a 25 year vision, Yearbook Med. Inform. 25 (S 01) (2016) S103–S116.
[15] Eta S. Berner, Tonya J. La Lande, Overview of clinical decision support systems, Clin. Decis. Support Syst. Theory Pract. (2016) 1–17.
[16] Tiffani J Bright, Anthony Wong, Ravi Dhurjati, Erin Bristow, Lori Bastian, Remy R Coeytaux, Gregory Samsa, Vic Hasselblad, John W Williams, Michael D Musty, et al., Effect of clinical decision-support systems: a systematic review, Ann. Internal Med. 157 (1) (2012) 29–43.
[17] Jonathan H. Chen, Tanya Podchiyska, Russ B. Altman, OrderRex: clinical order decision support and outcome predictions by data-mining electronic medical records, J. Am. Med. Inform. Assoc. 23 (2) (2016) 339–348.
[18] Wui Ip, Priya Prahalad, Jonathan Palma, Jonathan H Chen, et al., A data-driven algorithm to recommend initial clinical workup for outpatient specialty referral: Algorithm development and validation using electronic health record data and expert surveys, JMIR Med. Inform. 10 (3) (2022) e30104.

[19] Haley S Hunter-Zinck, Jordan S Peck, Tania D Strout, Stephan A Gaehde, Predicting emergency department orders with multilabel machine learning techniques and simulating effects on length of stay, J. Am. Med. Inform. Assoc. 26 (12) (2019) 1427–1436.
[20] Yiye Zhang, Rema Padman, James E. Levin, Paving the COWpath: data-driven design of pediatric order sets, J. Am. Med. Inform. Assoc. 21 (e2) (2014) e304–e311.
[21] Jeffrey Klann, Gunther Schadow, J. Michael McCoy, A recommendation algorithm for automating corollary order generation, in: AMIA Annual Symposium Proceedings, vol. 2009, American Medical Informatics Association, 2009, p. 333.
[22] Jonathan X Wang, Delaney K Sullivan, Alex C Wells, Jonathan H Chen, ClinicNet: Machine learning for personalized clinical order set recommendations, JAMIA Open 3 (2) (2020) 216–224.
[23] Morteza Noshad, Ivana Jankovic, Jonathan H. Chen, Clinical recommender algorithms to simulate digital specialty consultations, in: PACIFIC SYMPOSIUM on BIOCOMPUTING 2022, World Scientific, 2021, pp. 290–300.
[24] Sungjin Park, Seongsu Bae, Jiho Kim, Tackeun Kim, Edward Choi, Graph-text multi-modal pre-training for medical representation learning, in: Conference on Health, Inference, and Learning, PMLR, 2022, pp. 261–281.
[25] Edward Choi, Cao Xiao, Walter Stewart, Jimeng Sun, Mime: Multilevel medical embedding of electronic health records for predictive healthcare, Adv. Neural Inf. Process. Syst. 31 (2018).
[26] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart, Jimeng Sun, GRAM: graph-based attention model for healthcare representation learning, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017, pp. 787–795.
[27] Edward Choi, Zhen Xu, Yujia Li, Michael Dusenberry, Gerardo Flores, Emily Xue, Andrew Dai, Learning the graphical structure of electronic health records with graph convolutional transformer, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34 no. 01, 2020, pp. 606–613.
[28] Zheng Liu, Xiaohan Li, Hao Peng, Lifang He, S Yu Philip, Heterogeneous similarity graph neural network on electronic health records, in: 2020 IEEE International Conference on Big Data (Big Data), IEEE, 2020, pp. 1196–1205.
[29] Qimai Li, Zhichao Han, Xiao-Ming Wu, Deeper insights into graph convolutional networks for semi-supervised learning, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
[30] Marinka Zitnikv, Monica Agrawal, Jure Leskovec, Modeling polypharmacy side effects with graph convolutional networks, Bioinformatics 34 (13) (2018) i457–i466.
[31] Zheng Liu, Xiaohan Li, Zeyu You, Tao Yang, Wei Fan, Philip Yu, Medical triage chatbot diagnosis improvement via multi-relational hyperbolic graph neural network, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 1965–1969.
[32] Tong Wu, Yunlong Wang, Yue Wang, Emily Zhao, Yilian Yuan, Leveraging graph-based hierarchical medical entity embedding for healthcare applications, Sci. Rep. 11 (1) (2021) 1–13.
[33] Alina Vretinaris, Chuan Lei, Vasilis Efthymiou, Xiao Qin, Fatma Özcan, Medical entity disambiguation using graph neural networks, in: Proceedings of the 2021 International Conference on Management of Data, 2021, pp. 2310–2318.
[34] Michelle M. Li, Kexin Huang, Marinka Zitnik, Graph representation learning in biomedicine and healthcare, Nat. Biomed. Eng. (2022) 1–17.
[35] Haozhe Jia, Haoteng Tang, Guixiang Ma, Weidong Cai, Heng Huang, Liang Zhan, Yong Xia, A convolutional neural network with pixel-wise sparse graph reasoning for COVID-19 lesion segmentation in CT images, Comput. Biol. Med. 155 (2023) 106698.
[36] Xinyue Li, Rui Guo, Jing Lu, Tao Chen, Xiaohua Qian, Causality-driven graph neural network for early diagnosis of pancreatic cancer in non-contrast computerized tomography, IEEE Trans. Med. Imaging (2023).
[37] Siyi Tang, Amara Tariq, Jared A Dunnmon, Umesh Sharma, Praneetha Elugunti, Daniel L Rubin, Bhavik N Patel, Imon Banerjee, Predicting 30-day all-cause hospital readmission using multimodal spatiotemporal graph neural networks, IEEE J. Biomed. Health Inf. (2023).
[38] Xiangxiang Zeng, Xinqi Tu, Yuansheng Liu, Xiangzheng Fu, Yansen Su, Toward better drug discovery with knowledge graph, Curr. Opin. Struct. Biol. 72 (2022) 114–126.
[39] Petar Veličković, Everything is connected: Graph neural networks, Curr. Opin. Struct. Biol. 79 (2023) 102538.
[40] Clemens Isert, Kenneth Atz, Gisbert Schneider, Structure-based drug design with geometric deep learning, Curr. Opin. Struct. Biol. 79 (2023) 102548.
[41] Chengsheng Mao, Liang Yao, Yuan Luo, MedGCN: Medication recommendation and lab test imputation via graph convolutional networks, J. Biomed. Inform. 127 (2022) 104000.
[42] Will Hamilton, Zhitao Ying, Jure Leskovec, Deepsnap, Adv. Neural Inf. Process. Syst. 30 (2017).
[43] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, Yoshua Bengio, Graph Attention Networks, Int. Conf. Learn. Representations (2018) URL https://openreview.net/forum?id=rJXMpikCZ.
[44] Ted E Palen, David Price, Susan Shetterly, Kristin B Wallace, Comparing virtual consults to traditional consults using an electronic health record: an observational case–control study, BMC Med. Inform. Decis. Mak. 12 (1) (2012) 1–10.
[45] Clare Liddy, Margo S Rowan, Amir Afkham, Julie Maranger, Erin Keely, Building access to specialist care through e-consultation, Open Med. 7 (1) (2013) e1.
[46] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, Adv. Neural Inf. Process. Syst. 32 (2019).
[47] Matthias Fey, Jan Eric Lenssen, Fast graph representation learning with PyTorch Geometric, 2019, arXiv preprint arXiv:1903.02428.
[48] PyG, 2022, https://www.pyg.org/, (Accessed 31 September 2022).
[49] Jure Leskovec, Andrej Krevl, SNAP datasets: Stanford large network dataset collection, 2014.
[50] DeepSNAP, 2022, https://snap.stanford.edu/deepsnap/index.html, (Accessed 31 September 2022).