MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation
P. Otero et al. (Eds.)

doi:10.3233/SHTI220171

Developing a Siamese Network for UTIs Risk Prediction in Immobile Patients Undergoing Stroke

Zidu Xu[a,*], Chen Zhu[b,*], Yaowen Gu[a], Si Zheng[a], Xiangyu Sun[b], Jing Cao[b], Xinjuan Wu[b], Jiao Li[a]

[a]Institute of Medical Information / Library, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China,
[b]Department of Nursing, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China
*(*Equal contributor first authors)*

Abstract

Stroke patients tend to suffer from immobility, which increases the possibility of post-stroke complications. Urinary tract infections (UTIs) are one of the complications as an independent predictor of poor prognosis of stroke patients. However, the incidence of new UTIs onsets during hospitalization was rare in most datasets with a prevalence of 4%. This imbalanced data distribution sets obstacles to establishing an accurate prediction model. Our study aimed to develop an effective prediction model to identify UTIs risk in immobile stroke patients, and (2) to compare its prediction performance with traditional machine learning models. We tackled this problem by building a Siamese Network leveraging commonly used clinical features to identifying patients with UTIs risk. Model derivation and validation were based on a nationwide dataset including 3982 Chinese patients. Results showed that the Siamese Network performed better than traditional machine learning models in imbalanced datasets (Sensitivity: 0.810; AUC: 0.828).

Keywords:

Risk Prediction; Stroke; Immobility; Urinary Tract Infections; Siamese Network

Introduction

Stroke has come to China top3 leading cause of death with high risk of disability and mortality, and large burden of disease[22]. Stroke patients are affected by cerebral nerve damage, which increases the opportunity of being immobile, thus suffering from complications from immobility[6,21]. Urinary tract infections (UTIs) are one of the common complications after stroke only second to pulmonary infection, which not only prolongs the length of stay (LOS), but also aggravates the patient's suffering, posing threats to the safety and quality of medical care[22,23]. Our previous work has proved that immobile patients with a primary diagnosis of stroke are 2 times of risk probability as others to develop UTIs[22,25]. Therefore, assessing and identifying high-risk groups of UTIs has become an important task for the management of complications of immobile patients undergoing stroke.

Previous work has provided well documented facts for risk factors and highly susceptible populations for UTIs, such as the elder and female patients[19,25]. Nevertheless, the risk prediction for UTIs is just emerging. Researchers have included commonly used clinical, laboratory, and demographical features to develop UTIs prediction models in pediatric patients (AUC:0.79), surgical patients (AUC:0.72), etc. [3, 16]. However, these linear-based models are unable to capture the complex and non-linear relationships, and imbalanced data caused by the rarity of UTIs samples even bring challenges to the classification task, hence the existing prediction tools are not effective enough[7]. Machine learning (ML) methods have been widely applied to break down these barriers, and specifically, have succeeded in predicting several post-stroke complications such as predicting the prognosis of cardiovascular diseases, including complications prediction such as pneumonia and venous thromboembolism[2,14,15]. A Siamese Network is a class of deep neural network architectures that contain two or more identical subnetworks widely used in object detection[1,5,18], sequence embedding[12,24], and classification tasks [17]. Based on similarity calculations, all samples could be paired as the model input and mapped into new sample spaces thus balancing and enhancing the data. Experiments have shown that Siamese Network has good prediction performance in imbalanced data sets[20].

In this study, we designed a Siamese network (SimNet) to identify the immobile patients at high UTIs risk. We conducted the prediction task and compared the performance to 6 ML models to examine the strengths of SimNet in the risk prediction task, hoping to give some indications of quality and safety improvement in stroke patients care.

Methods

Study population

This study was based on the data from Common Complications of Bedridden Patients and the Construction of Standardized Nursing Intervention Model (CCBPC)[26], a nationwide investigation enrolling 23985 immobile patients from 25 hospitals throughout China. We enrolled 23985 immobile patients who are immobile for over 24 consecutive hours after admission (≥18 years old). 4018 patients with a primary diagnosis of stroke were then selected, of them 3982 patients were included to develop a prediction model identifying new UTIs onsets during hospitalization. This study was approved by the ethics review committee of Peking Union Medical College Hospital (S-700).

In this study, we defined stroke referring to the International Classification of Diseases, Tenth Revision (ICD-10) (I60~I64.x)[13] and UTIs to European Association of Urology Urologic Infections Guidelines[8]. Importantly, stroke patients in this study were categorized as hemorrhagic (I60.x, I61.x) stroke, ischemic (H34.1, I63.x, I64.x) stroke according to ICD-10, or have both subtypes of stroke above (Mixed cerebrovascular disease)[9,10]. Figure 1 shows the inclusion and exclusion criteria of this study.

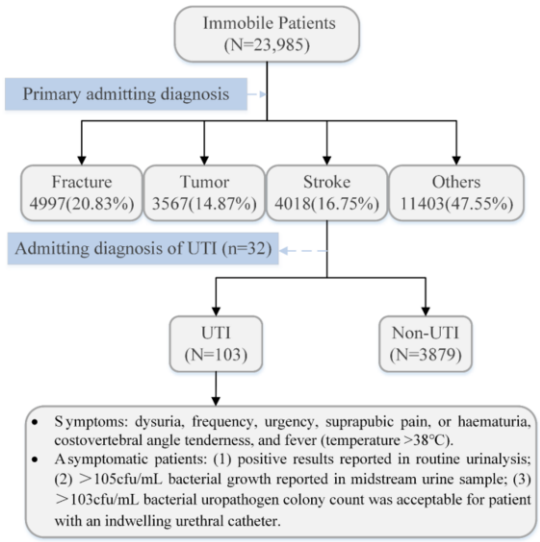

Figure 1-The patient flow diagram according to the inclusion and exclusion criteria.

Data preparation

A total of 18 features were included based on clinical knowledge. The entire dataset was split at the ratio of 6:2:2 with a stratified strategy to ensure the variables of interests, UTIs incidence particularly, shared similar data distributions. All models were assessed in test set (n=788), and the training (n=2388) and validation (n=788) sets were built for model derivation. After the data split, we performed data preprocessing such as missing value imputation and data transformation based on the features of the training set to avoid data leakage. The extracted features cover the (1) demographics; (2) hospitalization information; (3) medication status; and (4) disease related information. During data preprocessing, we set cutoffs of 1 week, 2 weeks for LOS, duration of immobility, and length of urethral catheter indwelling.

Model derivation and evaluation

The SimNet architecture for UTIs risk prediction is shown in Figure 2. SimNet is a two-tower structure, which mainly includes two shared parameter feature extraction blocks and a class prediction block. Specifically, the training process of SimNet includes the following stages:

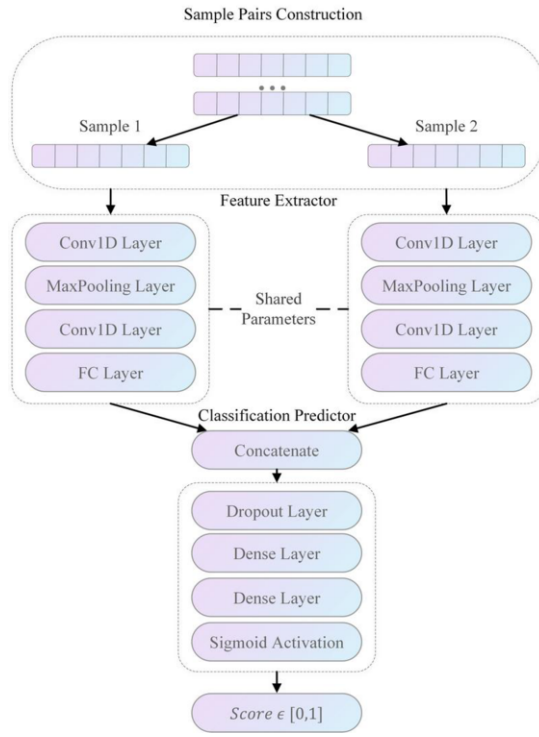

Figure 2–The Architecture of SimNet

Sample pairs construction. Given sample $\boldsymbol{x}$ and its true label $\boldsymbol{y}$, the sample set $\boldsymbol{X^{N\times D}}$, SimNet accepts the feature vectors of two samples as input, and pair each sample with a positive sample and a negative sample to construct a sample pair data set and use it as model input:

$$\boldsymbol{X}_{input} = \{(\boldsymbol{x}^1, \boldsymbol{x}^1_{pos}), (\boldsymbol{x}^1, \boldsymbol{x}^1_{neg}), \dots, (\boldsymbol{x}^n, \boldsymbol{x}^n_{pos}), (\boldsymbol{x}^n, \boldsymbol{x}^n_{neg})\}$$

$$\boldsymbol{x}^i_{pos} = \text{Random}(\boldsymbol{x}^k | \boldsymbol{y}^k = \boldsymbol{y}^i,\ \boldsymbol{x}^k \epsilon \boldsymbol{X})$$

$$\boldsymbol{x}^i_{neg} = \text{Random}(\boldsymbol{x}^k | \boldsymbol{y}^k \neq \boldsymbol{y}^i,\ \boldsymbol{x}^k \epsilon \boldsymbol{X})$$

Feature extraction. Extract hidden space embedding vectors $\boldsymbol{e}$ from two input samples respectively by the feature extraction blocks including 1D Convolution layer, MaxPooling layer, and Fully-Connected layer.

Classification. Concatenate the embedding vector $\boldsymbol{e}^i$,$\boldsymbol{e}^k$ of two samples, and transfer it as the input for category prediction block including Dropout and MLP layers to predict the probability that the two samples belong to the same category.

The details of the model parameters are shown in Table1. After the training process, we construct sample pairs between the test set samples and training set samples, and input the sample pairs into SimNet to get a probability score. Further, calculate the predicted probability of whether the test sample is UTIs based on the true label of the training set:

$$\boldsymbol{y}^{i,k}_{output} = \begin{cases} 1 - \boldsymbol{y}^{i,k}_{score} & if\ \boldsymbol{y}^k = 0 \\ \boldsymbol{y}^{i,k}_{score} & if\ \boldsymbol{y}^k = 1 \end{cases}$$

$$\boldsymbol{y}^i_{output} = \frac{1}{N} \sum_{k=1}^{k \epsilon N} \boldsymbol{y}^{i,k}_{output}$$

We examined the strengths of SimNet by making comparisons with linear models: logistic regression model with and without regularization, tree-based models: Random Forest(RF),

XGBoost (XGB), LightGBM (LGB), CatBoost (CTB), and multiple layer perception (MLP). To ameliorate the interference caused by imbalanced data, we used (1) a combination of over-sampling and undersampling - SMOTETomek; (2) an automatic hyper-parameter search tool based on neural architecture search (NAS) - NNI (Neural Network Intelligence) to improve the prediction effects of ML models. We use Sensitivity, Specificity, Accuracy, AUC, and Precision to evaluate all prediction models:

All procedures of data preprocessing, model derivation, and validation were conducted in Python 3.8. The model algorithms and hyperparameters tuning tool were based on Python Library: sci-kit learn, Tensorflow and NNI.

Table 1-SimNet Parameters

Parameters			
Filter	Kernel	Activation	Loss
64	(1,3)	LeakyReLU	Cross Entropy
Optimizer	Batch Size	Learning Rate	Epochs
Adam	512	1e-4	400

Results

Characteristics of subjects

Of 3982 patients included in this study, 103(2.59%) were diagnosed with new UTIs onsets during hospitalization. Data of outcomes and features distribution between training, validation, and test sets. Significant statistically difference was report between patients with and without UTIs in gender distribution (Table 2), hospitalization features (i.e, LOS, experience of surgery and duration of immobility), disease related features such as the proportion of classifications of stroke, pneumonia, disturbance of consciousness, medical status such as during of catheterization, glucocorticoid use, perineal care, urethral invasion procedures, urinary catheter change, and urinary drainage bag change (within 1 month).

Prediction Model Performances

The performances of logistic regression(LR), five ML-based model(RLR, RF, LGB, CTB, MLP) and a deep learning-based model(SimNet) are shown in Table 3. The results show that the prediction performance of SimNet on the five metrics is significantly better than other models. Compared with the sub-optimal results, SimNet achieved 4.5%, 0.7%, 1.1%, 3.0%, and 14.7% improvement in Sensitivity, Specificity, Accuracy, AUC, and Precision, respectively.

Table 2-Statistical Description of Baseline Characteristics

Categories	Variable names	Description	Mean(SD)/ Counts (Percentage)
Outcome (n=1)	UTIs		103(2.59%)
Demographics (n=2)	Gender	Male	2316(58.16%)
	Age, y		63.79(14.76)
Hospitalization information (n=3)	LOS, d	1~7 d	612(15.37%)
		8~14 d	1580(39.63%)
		≥15 d	1790(45.00%)
	Duration of immobility	1~7 d	2124(53.34%)
		8~14 d	977(24.54%)
		≥15 d	881(22.12%)
	Experience of surgery		878(22.00%)
Disease related information (n=6)	Classification of stroke	Ischaemic	2383(59.84%)
		Haemorrhagic	1396(35.06%)
		Mixed	203(5.10%)
	Pneumonia		883(22.17%)
	Diabetes mellitus		523(13.13%)
	Disturbance of consciousness		1367(34.33%)
	Urinary incontinence		390(9.79%)
	Serum albumin		37.78(6.43)
Medication status (n=7)	Invasive mechanical ventilation		611(15.33%)
	Urethral invasion procedures (within 1 month)		1196(30.00%)
	Duration of catheterization	0 d	2030(50.98%)
		1~7 d	1044(26.22%)
		8~14 d	420(10.55%)
		>14 d	488(12.26%)
	Glucocorticoid use		567(14.24%)
	Perineal care		2461(61.80%)
	Urinary catheter change (within 1 month)		219(11.33%)
	Urinary drainage bag change (within 1 month)		891(46.09%)

Among the traditional linear regression and five machine learning models, RLR achieved a sub-optimal sensitivity but got the worst specificity. MLP achieves sub-optimal results in four metrics (Specificity, Accuracy, AUC, Precision), but the ability to detect patients with UTI risk was undesirable, representing a sensitivity of 71.4%.

Table 3– Prediction Model Performances

Prediction Model	Performance Metrics				
	Sensitivity	Specificity	Accuracy	AUC	Precision
LR	76.2	68.9	69.1	79.2	6.2
RLR	77.5	64.7	65.0	80.2	5.6
RF	74.0	68.7	68.9	79.0	6.0
LGB	76.2	69.3	69.5	78.8	6.3
CTB	66.7	69.7	69.6	80.3	5.6
MLP	71.4	73.5	73.4	80.4	6.8
SimNet	**81.0**	**74.0**	**74.2**	**82.8**	**7.8**

Discussion

In this study, we proved that SimNet gives a good performance in predicting UTIs risk for immobile stroke patients. The improvement in sensitivity metrics is clinically meaningful especially for the outcome of interests of the minority class, where the ability to identify potential high-risk patients is of great significance. Compared to traditional ML models with a resampling strategy, SimNet performs the best in all performance metrics and achieves a balance in sensitivity and specificity. Therefore, we may suppose this accurate classifier to have the potential for clinical practice. To summarize, our results showed that the SimNet can be an effective prediction model of clinical significance to help improve identifying infrequent UTIs incidence in immobile stroke patients.

One reason attribute to the better performance of SimNet might be the in-pair inputs of samples, which could contribute to a meaningful vector representation during the training procedure, thus conducting the classification task easier. Figure 3 shows that in contrast with training sample space, the embedding vector of SimNet feature extraction block narrows the distance between UTIs samples. Furthermore, the ability to reuse samples in the training set might also help enhance the prediction effect.

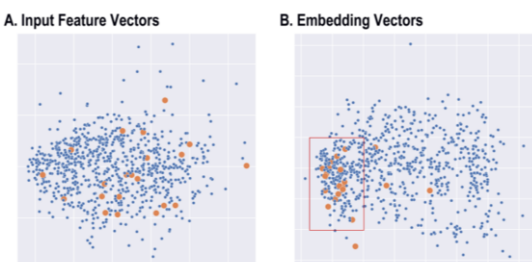

Figure 3–Vector Visualization Result by PCA

We acknowledge several limits in this study. First, for the lack of an external validation set, we are unsure whether the prediction effect of the SimNet is robust. Further investigations including larger sample size and wider geographical range remain needed to refine the generalizability and usability of this prediction model. It has been accepted that data augmentation is critical for the detection of infrequent or even rare incidence in medical settings. Hence the future direction might step into the efficient use of limited labeled data or massive unlabeled data, thus make the clinical risk identification more 'intelligent'[4,11].

Conclusions

In this study, we aimed to predict the infrequent UTIs onsets during hospitalization among stroke patients during immobility. We built a large population-based cohort for prediction model derivation and evaluation. A deep learning-based classifier with a Siamese Network architecture was proposed. Meanwhile, we trained 6 machine learning models in the same cohort with a resampling strategy for optimal prediction. We conducted comparisons in 5 evaluation metrics and proved that the Siamese Network had better performance than other machine learning models in our target issue with off-balance data. We partly attributed this strength to the characteristic of in-pair inputs, which helped the augmentation of minority class to optimize the prediction effects.

Acknowledgements

This study was supported by Beijing Natural Sciences Grants (Z200016) and National Health Commission of the People's Republic of China (grant number 201502017). All the authors gratefully acknowledge the following 6 nursing directors for their excellent work in data collection: Baoyun Song from Henan Provincial People's Hospital, Zhengzhou, China; Jingfen Jin from The Second Affiliated Hospital Zhejiang University School of Medicine, Hangzhou, China; Yilan Liu from Union Hospital of Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China; Xianxiu Wen from Sichuan Provincial People's Hospital, Chengdu, China; Shouzhen Cheng from The First Affiliated Hospital, Sun Yat-sen University, Guangzhou, China.

References

[1] M.H. Abdelpakey and M.S. Shehata, DP-Siam: Dynamic Policy Siamese Network for Robust Object Tracking, *IEEE Trans Image Process* (2019).

[2] B. Ambale-Venkatesh, X. Yang, C.O. Wu, K. Liu, W.G. Hundley, R. McClelland, A.S. Gomes, A.R. Folsom, S. Shea, E. Guallar, D.A. Bluemke, and J.A.C. Lima, Cardiovascular Event Prediction by Machine Learning: The Multi-Ethnic Study of Atherosclerosis, *Circulation research* **121** (2017), 1092-1101.

[3] K.Y. Bilimoria, Y. Liu, J.L. Paruch, L. Zhou, T.E. Kmiecik, C.Y. Ko, and M.E. Cohen, Development and evaluation of the universal ACS NSQIP surgical risk calculator: a decision aid and informed consent tool for patients and surgeons, *J Am Coll Surg* **217** (2013), 833-842.e831-833.

[4] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, A simple framework for contrastive learning of visual representations, in: *International conference on machine learning*, PMLR, 2020, pp. 1597-1607.

[5] M. Dunnhofer, M. Antico, F. Sasazawa, Y. Takeda, S. Camps, N. Martinel, C. Micheloni, G. Carneiro, and D. Fontanarosa, Siam-U-Net: encoder-decoder siamese network for knee cartilage tracking in ultrasound images, *Med Image Anal* **60** (2020), 101631.

[6] H.C.A. Emsley and S.J. Hopkins, Acute ischaemic stroke and infection: recent and emerging concepts, *The Lancet. Neurology* **7** (2008), 341-353.

[7] B.A. Goldstein, A.M. Navar, and R.E. Carter, Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges, *Eur Heart J* **38** (2017), 1805-1814.

[8] J. Groen, J. Pannek, D. Castro Diaz, G. Del Popolo, T. Gross, R. Hamid, G. Karsenty, T.M. Kessler, M. Schneider, L. t Hoen, and B. Blok, Summary of European Association of Urology (EAU) Guidelines on Neuro-Urology, *European urology* **69** (2016), 324-333.

[9] G.J. Hankey, Stroke, *The Lancet* **389** (2017), 641-654.

[10] D. He, Y. Yu, S. Wu, S. Tian, H. Yu, S. Xu, and L. Chu, Mixed cerebrovascular disease in an elderly patient with mixed vascular risk factors: a case report, *BMC Neurology* **19** (2019), 26.

[11] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, Momentum contrast for unsupervised visual representation learning, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729-9738.

[12] M. Jeon, D. Park, J. Lee, H. Jeon, M. Ko, S. Kim, Y. Choi, A.C. Tan, and J. Kang, ReSimNet: drug response similarity prediction using Siamese neural networks, *Bioinformatics* **35** (2019), 5249-5256.

[13] R.A. Kokotailo and M.D. Hill, Coding of stroke and stroke risk factors using international classification of diseases, revisions 9 and 10, *Stroke* **36** (2005), 1776-1781.

[14] S. Kumar, M.H. Selim, and L.R. Caplan, Medical complications after stroke, *The Lancet. Neurology* **9** (2010), 105-118.

[15] W. Liang, H. Liang, L. Ou, B. Chen, A. Chen, C. Li, Y. Li, W. Guan, L. Sang, J. Lu, Y. Xu, G. Chen, H. Guo, J. Guo, Z. Chen, Y. Zhao, S. Li, N. Zhang, N. Zhong, and J. He, Development and Validation of a Clinical Risk Score to Predict the Occurrence of Critical Illness in Hospitalized Patients With COVID-19, *JAMA internal medicine* **180** (2020), 1081-1089.

[16] R. Luciano, S. Piga, L. Federico, M. Argentieri, F. Fina, M. Cuttini, E. Misirocchi, F. Emma, and M. Muraca, Development of a score based on urinalysis to improve the management of urinary tract infection in children, *Clinica Chimica Acta* **413** (2012), 478-482.

[17] S.A. Memon, K.A. Khan, and H. Naveed, HECNet: a hierarchical approach to enzyme function classification using a Siamese Triplet Network, *Bioinformatics* **36** (2020), 4583-4589.

[18] J. Shen, X. Tang, X. Dong, and L. Shao, Visual Object Tracking by Hierarchical Attention Siamese Network, *IEEE Trans Cybern* **50** (2020), 3068-3080.

[19] C. Smith, E. Almallouhi, and W. Feng, Urinary tract infection after stroke: A narrative review, *Journal of the neurological sciences* **403** (2019), 146-152.

[20] D. Sun, Z. Wu, Y. Wang, Q. Lv, and B. Hu, Risk Prediction for Imbalanced Data in Cyber Security : A Siamese Network-based Deep Learning Classification Framework, in: *2019 International Joint Conference on Neural Networks (IJCNN)*, 2019, pp. 1-8.

[21] W. Wang, B. Jiang, H. Sun, X. Ru, D. Sun, L. Wang, L. Wang, Y. Jiang, Y. Li, Y. Wang, Z. Chen, S. Wu, Y. Zhang, D. Wang, Y. Wang, and V.L. Feigin, Prevalence, Incidence, and Mortality of Stroke in China: Results from a Nationwide Population-Based Survey of 480 687 Adults, *Circulation* **135** (2017), 759-771.

[22] Y.-J. Wang, Z.-X. Li, H.-Q. Gu, Y. Zhai, Y. Jiang, X.-Q. Zhao, Y.-L. Wang, X. Yang, C.-J. Wang, X. Meng, H. Li, L.-P. Liu, J. Jing, J. Wu, A.-D. Xu, Q. Dong, D. Wang, and J.-Z. Zhao, China Stroke Statistics 2019: A Report From the National Center for Healthcare Quality Management in Neurological Diseases, China National Clinical Research Center for Neurological Diseases, the Chinese Stroke Association, National Center for Chronic and Non-communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention and Institute for Global Neuroscience and Stroke Collaborations, **5** (2020), 211-239.

[23] S. Wu, B. Wu, M. Liu, Z. Chen, W. Wang, C.S. Anderson, P. Sandercock, Y. Wang, Y. Huang, L. Cui, C. Pu, J. Jia, T. Zhang, X. Liu, S. Zhang, P. Xie, D. Fan, X. Ji, K.-S.L. Wong, and L. Wang, Stroke in China: advances and challenges in epidemiology, prevention, and management, *The Lancet. Neurology* **18** (2019), 394-405.

[24] W. Zheng, L. Yang, R.J. Genco, J. Wactawski-Wende, M. Buck, and Y. Sun, SENSE: Siamese neural network for sequence embedding and alignment-free comparison, *Bioinformatics* **35** (2019), 1820-1828.

[25] C. Zhu, H. Liu, Y. Wang, J. Jiao, Z. Li, J. Cao, B. Song, J. Jin, Y. Liu, X. Wen, S. Cheng, and X. Wu, Prevalence, incidence, and risk factors of urinary tract infection among immobile inpatients in China: a prospective, multi-centre study, *The Journal of hospital infection* **104** (2020), 538-544.

[26] C. Zhu, H. Liu, Y. Wang, J. Jiao, Z. Li, J. Cao, B. Song, J. Jin, Y. Liu, X. Wen, S. Cheng, and X. Wu, Prevalence, incidence, and risk factors of urinary tract infection among immobile inpatients in China: a prospective, multi-centre study, *J Hosp Infect* **104** (2020), 538-544.

Address for correspondence

Corresponding author: Jiao Li, PhD, Institute of Medical Information / Medical Library, Chinese Academy of Medical Sciences & Peking Union Medical College, No. 3 Yabao Road, Chaoyang District, Beijing, 100020, China, Email: li.jiao@imicams.ac.cn.