

Prediction of antibiotic-associated cutaneous adverse drug reactions using electronic health record foundation models

Received: 7 May 2025

Accepted: 19 February 2026

Cite this article as: Kim, J., Kim, K., Yun, J.-E. *et al.* Prediction of antibiotic-associated cutaneous adverse drug reactions using electronic health record foundation models. *npj Digit. Med.* (2026). <https://doi.org/10.1038/s41746-026-02503-x>

Junmo Kim, Kyunghoon Kim, Jeong-Eun Yun, Yu-Kyoung Hwang, Min-Gyu Kang, Seok Kim, Sooyoung Yoo, Chaiho Shin, Suhyun Kim, Kwangsoo Kim & Sae-Hoon Kim

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

Prediction of Antibiotic-Associated Cutaneous Adverse Drug Reactions Using Electronic Health Record Foundation Models

Junmo Kim^{1,*}, Kyunghoon Kim^{2,3,*}, Jeong-Eun Yun⁴, Yu-Kyoung Hwang⁵, Min-Gyu Kang⁵, Seok Kim⁶, Sooyoung Yoo⁶, Chaiho Shin⁷, Suhyun Kim^{8,9}, Kwangsoo Kim^{8,10,11,†}, Sae-Hoon Kim^{12,13,14,†}

¹Interdisciplinary Program in Bioengineering, Seoul National University, Seoul, Republic of Korea

²Department of Pediatrics, Seoul National University Bundang Hospital, Seongnam, Republic of Korea

³Department of Pediatrics, Seoul National University College of Medicine, Seoul, Republic of Korea

⁴Department of Internal Medicine, Chung-Ang University College of Medicine, Seoul, Republic of Korea

⁵Department of Internal Medicine, College of Medicine, Chungbuk National University and Chungbuk National University Hospital, Cheongju, Republic of Korea

⁶Healthcare ICT Research Center, Office of eHealth Research and Businesses, Seoul National University Bundang Hospital, Seongnam, Republic of Korea

⁷Interdisciplinary Program of Medical Informatics, Seoul National University, Seoul, Republic of Korea

⁸Department of Transdisciplinary Medicine, Institute of Convergence Medicine with Innovative Technology, Seoul National University Hospital

⁹Department of Clinical Medical Sciences, College of Medicine, Seoul National University, Seoul, Republic of Korea

¹⁰Center for Data Science, Healthcare AI Research Institute, Seoul National University Hospital

¹¹Department of Medicine, College of Medicine, Seoul National University

¹²Department of Internal Medicine, Seoul National University Bundang Hospital, Seongnam, Republic of Korea

¹³Department of Internal Medicine, Seoul National College of Medicine, Seoul, Republic of Korea

¹⁴Institute of Allergy and Clinical Immunology, Seoul National University Medical Research Center Seoul, Republic of Korea

*These authors contributed equally.

† Corresponding Authors**Kwangsoo Kim, Ph.D.**

Department of Transdisciplinary Medicine, Institute of Convergence Medicine with Innovative Technology, Seoul
National University Hospital

Address: 101, Daehak-ro, Jongno-gu, Seoul, Republic of Korea

Email: kwangsookim@snu.ac.kr

Sae-Hoon Kim, M.D., Ph.D.

Department of Internal Medicine, Seoul National University Bundang Hospital, Seongnam, Republic of Korea

Address: 82, Gumi-ro 173beon-gil, Bundang-gu, Seongnam-si, Gyeonggi-do, Republic of Korea

E-mail: shkrins@snu.ac.kr

Abstract

Cutaneous adverse drug reactions (CADRs) are the most common form of adverse drug reactions, ranging from mild rashes to life-threatening diseases, such as Stevens–Johnson syndrome and toxic epidermal necrolysis. However, there is no effective tool to predict antibiotic-associated CADRs. In this study, we propose an antibiotic-associated CADR prediction model using electronic health record (EHR) foundation models (FMs). EHR FMs are based on the pretraining-finetuning paradigms of language models, corresponding medical codes and their sequences to words and sentences. We included 802,131 inpatients across three tertiary hospitals in Korea, combining EHR data with nursing statements and reports to extract skin rash records. Our approach achieved the best predictive performance compared to all the other baseline models across all datasets. To enhance clinical relevance, we classified CADRs into immediate and delayed types and conducted a detailed sub-analysis. Finally, we found that properly configured EHR FMs can effectively predict the risk of developing antibiotics-associated CADRs, particularly for delayed-type reactions where predictive testing options are limited.

Introduction

Adverse drug reaction (ADR) is a harmful or unpleasant response to a medicinal product that predicts hazard for future administration, requiring prevention, specific treatment, dose adjustment, or product withdrawal.¹ ADRs are among the leading causes of death in the United States, and serious ADRs occur in approximately 6.7% of hospitalized patients, with fatal ADRs reported in 0.32% of cases.² In a prospective study of 5,000 patients hospitalized at three tertiary hospitals in Korea, ADRs were observed in 10.2% of cases, and those who developed ADRs had a 5-day longer hospital stay and significantly higher medical expenses compared to the control group.³ ADRs worsen clinical outcomes and increase medical costs because of prolonged hospital stays and additional visits³⁻⁵.

Among the various manifestations, skin is most commonly affected by ADRs, involving mild rashes to life-threatening diseases, such as Stevens–Johnson syndrome (SJS) and toxic epidermal necrolysis (TEN). Antibiotics are among the most common triggers for cutaneous adverse drug reactions (CADRs).^{6,7} In a prospective study from an Italian university hospital, ADRs occurred in 4% of cases, of which 28% were CADRs.⁸ In a French general hospital, the prevalence of CADR was 3.6 per 1000 patients, 34% of whom were classified as severe, with antibiotics, particularly penicillin, being the most commonly implicated drug.⁹ Serious CADRs are rare but potentially life-threatening drug hypersensitivity reactions including Drug Rash with Eosinophilia and Systemic Symptoms (DRESS) syndrome, SJS, and TEN, which have reported mortality rate of 20% to 40%.^{10,11} Thus, predicting CADRs and preventing deterioration is important; however, a consistent and standardized method for predicting CADRs has not yet been established.

The adoption of electronic health records (EHRs) has grown significantly worldwide, leading to a substantial increase in healthcare data¹²⁻¹⁴. Subsequently, deep learning-based EHR analysis models have been developed using the large-scale patient data, and representations of patient trajectories have emerged as a promising area of research¹⁵⁻²⁰. As individual medical codes and sequential sets of medical codes correspond to words and sentences, several studies have proposed EHR foundation models (FMs) based on the pretraining paradigms used in language models (LMs). For pretraining, although most studies have utilized the masked LM²¹, autoregressive LMs have also been used²². Notably, previous studies have achieved excellent performance in predicting several diseases, such as pancreatic cancer¹⁸, heart failure¹⁹, non-accidental trauma¹⁵, and various ADRs²³ using EHR FMs and patient trajectories.

The present study aims to develop a CADR prediction model using EHR FMs and patient trajectories

consisting of diagnosis, measurement, prescription, and procedure records from EHR. We compared the performance of several machine learning, deep learning, and EHR foundation models, including BEHRT¹⁷, CLMBR-t-base²², and CDM-BERT²³. Specifically, we demonstrated the effectiveness of domain embedding (DE) in CDM-BERT, which exhibited superior performance in predicting various ADRs. Skin rash records were extracted from nursing records (including statements and reports) and integrated with the EHR data. For internal and external validation, we utilize data from three tertiary hospitals in the Republic of Korea (hereafter, Korea). Unlike prior research, this study demonstrates true generalizability with stable performance across institutions without additional finetuning on external datasets. Moreover, we classify CADR into immediate and delayed types for a detailed sub-analysis, and interpret the results based on the clinical background. With a rigorously defined practical cohort for CADR prediction, this study represents a progression in ADR prediction from proof-of-concept to clinically implementable solutions.

Results

Study population

We used the records of adult patients who were hospitalized for at least 3 days and were prescribed antibiotics during the stay. Since we used nursing statements and reports for CADR record extraction, only the data from inpatients were used because accurate labeling of antibiotic-associated CADR in outpatients was not feasible due to the lack of nursing records. Patients from three tertiary hospitals in Korea—Seoul National University Hospital (SNUH; n = 366,434), Seoul National University Bundang Hospital (SNUBH; n = 325,396), and Chungbuk National University Hospital (CNUH; n = 110,301)—were included in this study. The prediction for all patients who were prescribed antibiotics was performed during their last visit. The exclusion criteria for patients with antibiotic-associated skin rash were as follows: <20 records, absence of skin rash during antibiotic treatment, and continuation of antibiotic treatment after the occurrence of skin rash. Given the need for continuous antibiotic treatment, patients who were prescribed medication for skin rash within 1 day after the occurrence were included, even if they were treated with antibiotics continuously. Finally, 1,906, 3,070, and 524 patients with CADR and 216,649, 311,631, and 34,357 patients without CADR from SNUH, SNUBH, and CNUH, respectively, were included in the study cohort (Figure 1). The antibiotics and the drugs used to treat skin rash are summarized in Supplementary Table 1.

Data from SNUH were used for training and internal validation of the CADR prediction model and data

from SNUBH and CNUH were used for external validation (Table 1). The incidence ratio of CADR was approximately 1%, consistent with that reported in previous studies²⁴⁻²⁶. The incidence ratio of CADR for each antibiotic is summarized in Supplementary Table 2. No significant differences were observed for age and sex between the CADR and non-CADR groups. However, the days of hospitalization and the number of codes per patient were higher in the CADR groups than in the non-CADR groups for all hospitals. The prevalence rates of all comorbidities were higher in the CADR group of SNUH; however, no evident pattern of comorbidities was observed for SNUBH and CNUH. The overall distribution of data varied between hospitals, and we demonstrated the validity of our model using these external datasets. For a more detailed analysis, we classified CADR into immediate and delayed types. The baseline characteristics of patients with immediate and delayed CADR types are summarized in Supplementary Table 3.

Table 1. Baseline characteristics of the included patients

	Internal dataset SNUH (n=218,555)			External dataset 1 SNUBH (n=314,701)			External dataset 2 CNUH (n=34,881)		
	CADR	Non-CADR	P-value	CADR	Non-CADR	P-value	CADR	Non-CADR	P-value
n	1,906 (0.87%)	216,649 (99.13%)	n/a	3,070 (0.98%)	311,631 (99.02%)	n/a	524 (1.50%)	34,357 (98.50%)	n/a
Mean age (SD) ¹	60.46 (16.11)	57.71 (16.21)	<0.0001	56.58 (16.59)	56.39 (16.87)	0.5345	60.40 (17.59)	61.05 (17.72)	0.4046
Male sex ²	51.42%	47.49%	0.0006	49.19%	45.22%	<0.0001	54.20%	53.92%	0.8970
Median days of hospitalization (Q1, Q3) ³	43 (12, 116)	20 (8, 63)	<0.0001	19 (8, 58)	17 (7, 51)	<0.0001	15 (5, 52)	12 (4, 39)	<0.0001
Median number of codes per patient (Q1, Q3) ³	772 (239, 1944)	255 (107, 785)	<0.0001	187 (82, 575)	136 (58, 405)	<0.0001	86 (38, 240)	67 (28, 175)	<0.0001
Comorbidities ²									
Myocardial infarction	0.84%	0.76%	0.7884	1.07%	1.06%	0.9967	1.72%	1.93%	0.8534
Congestive heart failure	4.93%	1.81%	<0.0001	2.74%	2.08%	0.0134	3.63%	3.45%	0.9142
Peripheral vascular disease	2.94%	1.44%	<0.0001	1.73%	1.15%	0.0038	0.76%	0.89%	0.5984
Cerebrovascular disease	8.03%	5.55%	<0.0001	7.65%	6.03%	0.0003	7.44%	6.68%	0.6571
Dementia	5.77%	3.27%	<0.0001	3.58%	3.57%	0.9956	2.67%	1.79%	0.3042
Pulmonary disease	6.35%	4.94%	0.0055	5.80%	6.17%	0.4179	2.67%	2.97%	0.6051
Connective tissue disease	3.52%	1.19%	<0.0001	2.15%	1.70%	0.0643	0.95%	0.64%	0.9312
Peptic ulcer disease	2.99%	2.06%	0.0091	1.07%	1.50%	0.0453	0.76%	0.44%	0.8991
Mild liver disease	6.72%	4.57%	<0.0001	3.19%	2.88%	0.3835	1.34%	0.72%	0.1631
Liver disease	1.15%	0.41%	<0.0001	0.29%	0.22%	0.7741	0.19%	0.41%	0.2651
Uncomplicated diabetes	12.59%	8.60%	<0.0001	6.38%	7.92%	0.0015	4.58%	5.27%	0.4313
Complicated diabetes	2.94%	1.36%	<0.0001	3.68%	4.76%	0.0046	3.82%	4.46%	0.5537
Paraplegia and hemiplegia	0.37%	0.22%	0.2618	0.85%	0.54%	0.0290	0.19%	0.39%	0.2854
Renal disease	8.24%	4.63%	<0.0001	3.71%	3.10%	0.0722	5.34%	4.67%	0.6698
Malignant tumor	37.41%	32.09%	<0.0001	21.11%	19.86%	0.0872	9.54%	11.68%	0.1176

*Abbreviations: SNUH=Seoul National University Hospital, SNUBH=Seoul National University Bundang Hospital, CNUH=Chungbuk National University Hospital, CADR=cutaneous adverse drug reaction, SD=standard deviation

¹P-values were calculated by the Student's t-test.

²P-values were calculated by the Mann-Whitney U test.

³P-values were calculated by the Fisher's exact test.

Prediction performance

For training, the prediction timepoint was set as the time of the latest antibiotic treatment before CADR. The models were designed to predict whether CADR would occur if the patient was prescribed specific antibiotics. The last prescribed antibiotics and the subsequent past records were used as inputs for the model. We evaluated the CADR prediction performance using eight models: two tree-based models (random forest²⁷ and gradient boosting machine (GBM)²⁸), three recurrent neural network (RNN)-based models (long short-term memory (LSTM)²⁹, gated recurrent unit (GRU)³⁰, and RETAIN³¹), and three EHR FMs (BEHRT¹⁷, CLMBR-t-base²², and CDM-BERT²³). We compared the performance of BEHRT and CDM-BERT including the models trained without pretraining. As sequence data (patient trajectory) cannot be used as input for the random forest and GBM, we transformed each sequence into a count-based one-dimensional vector³². For the remaining models, we used the sequence data with age and segment embeddings. We randomly split the patient data from SNUH into development (80%) and hold-out internal validation (20%) datasets. CDM-BERT demonstrated the best prediction performance for internal and external validations, with area under the receiver operating characteristic (ROC) curve (AUROC) values of 0.975, 0.928, and 0.893 for SNUH, SNUBH, and CNUH, respectively (Table 2). Notably, the performance was maintained even without the additional fine-tuning for external validation. Meanwhile, BEHRT and CLMBR-t-base poorly predicted the CADR, even after pretraining. Despite the low incidence ratio of CADR (approximately 1%), CDM-BERT achieved high precision of 15.8%, 7.8%, and 12.6% for SNUH, SNUBH, and CNUH, respectively. The ROC and precision-recall (PR) curves of all models are summarized in Supplementary Figure 1.

Table 2. Performance for predicting CADR.

	AUROC	AUPRC	Sensitivity	Specificity	Precision	F1-score
Internal dataset (SNUH)						
Random forest	0.747 (0.723-0.771)	0.095 (0.070-0.119)	0.647 (0.643-0.652)	0.704 (0.699-0.708)	0.020 (0.018-0.021)	0.038 (0.036-0.040)
Gradient boosting machine	0.782 (0.759-0.805)	0.065 (0.041-0.088)	0.627 (0.623-0.632)	0.783 (0.779-0.787)	0.026 (0.024-0.027)	0.049 (0.047-0.052)
LSTM	0.606 (0.572-0.639)	0.019 (0.000-0.639)	0.514 (0.509-0.519)	0.719 (0.715-0.723)	0.016 (0.015-0.018)	0.032 (0.030-0.034)
GRU	0.682 (0.656-0.707)	0.024 (0.000-0.707)	0.743 (0.739-0.747)	0.541 (0.537-0.546)	0.015 (0.014-0.016)	0.029 (0.027-0.030)
RETAIN	0.751 (0.727-0.774)	0.032 (0.009-0.056)	0.665 (0.661-0.669)	0.715 (0.711-0.719)	0.021 (0.020-0.022)	0.041 (0.039-0.042)
BEHRT (No pretraining)	0.781 (0.760-0.802)	0.032 (0.011-0.053)	0.756 (0.752-0.760)	0.686 (0.682-0.690)	0.022 (0.020-0.023)	0.042 (0.040-0.044)

BEHRT (Pretrained)	0.770 (0.748-0.791)	0.034 (0.013-0.056)	0.831 (0.828-0.835)	0.569 (0.564-0.573)	0.017 (0.016-0.019)	0.034 (0.032-0.036)
CLMBR-t-base (Pretrained)	0.772 (0.751-0.794)	0.036 (0.014-0.058)	0.695 (0.691-0.700)	0.711 (0.706-0.715)	0.022 (0.020-0.023)	0.042 (0.040-0.044)
CDM-BERT (No pretraining)	0.897 (0.880-0.914)	0.124 (0.107-0.141)	0.806 (0.802-0.810)	0.861 (0.858-0.864)	0.050 (0.048-0.052)	0.095 (0.092-0.098)
CDM-BERT (Pretrained)	0.975 (0.967-0.983)	0.309 (0.301-0.317)	0.967 (0.966-0.969)	0.953 (0.951-0.955)	0.158 (0.155-0.162)	0.272 (0.268-0.276)
External dataset 1 (SNUBH)						
Random forest	0.625 (0.615-0.636)	0.025 (0.014-0.035)	0.613 (0.611-0.615)	0.582 (0.580-0.584)	0.014 (0.014-0.015)	0.028 (0.027-0.028)
Gradient boosting machine	0.588 (0.577-0.599)	0.019 (0.008-0.029)	0.483 (0.481-0.484)	0.670 (0.669-0.672)	0.014 (0.014-0.015)	0.028 (0.027-0.028)
LSTM	0.545 (0.533-0.556)	0.013 (0.002-0.024)	0.528 (0.527-0.530)	0.583 (0.581-0.585)	0.012 (0.012-0.013)	0.024 (0.024-0.025)
GRU	0.599 (0.589-0.609)	0.014 (0.004-0.023)	0.566 (0.564-0.568)	0.588 (0.586-0.589)	0.013 (0.013-0.014)	0.026 (0.026-0.027)
RETAIN	0.434 (0.424-0.445)	0.009 (0.000-0.445)	0.023 (0.022-0.023)	0.984 (0.983-0.984)	0.013 (0.013-0.014)	0.017 (0.016-0.017)
BEHRT (No pretraining)	0.592 (0.582-0.601)	0.013 (0.004-0.023)	0.694 (0.692-0.695)	0.452 (0.451-0.454)	0.012 (0.012-0.013)	0.024 (0.024-0.025)
BEHRT (Pretrained)	0.603 (0.593-0.614)	0.016 (0.006-0.026)	0.550 (0.548-0.552)	0.604 (0.603-0.606)	0.014 (0.013-0.014)	0.026 (0.026-0.027)
CLMBR-t-base (Pretrained)	0.600 (0.590-0.610)	0.015 (0.005-0.025)	0.621 (0.619-0.622)	0.520 (0.519-0.522)	0.013 (0.012-0.013)	0.025 (0.024-0.025)
CDM-BERT (No pretraining)	0.850 (0.842-0.858)	0.062 (0.054-0.069)	0.762 (0.760-0.763)	0.837 (0.835-0.838)	0.044 (0.043-0.045)	0.083 (0.082-0.084)
CDM-BERT (Pretrained)	0.928 (0.923-0.933)	0.187 (0.182-0.192)	0.882 (0.881-0.884)	0.897 (0.896-0.898)	0.078 (0.077-0.079)	0.144 (0.142-0.145)
External dataset 2 (CNUH)						
Random forest	0.553 (0.528-0.577)	0.024 (0.000-0.577)	0.471 (0.466-0.477)	0.638 (0.633-0.643)	0.019 (0.018-0.021)	0.037 (0.035-0.039)
Gradient boosting machine	0.552 (0.527-0.577)	0.018 (0.000-0.577)	0.525 (0.520-0.530)	0.587 (0.582-0.592)	0.019 (0.018-0.021)	0.037 (0.035-0.039)
LSTM	0.539 (0.512-0.566)	0.020 (0.000-0.566)	0.395 (0.390-0.400)	0.725 (0.720-0.730)	0.021 (0.020-0.023)	0.041 (0.039-0.043)
GRU	0.600 (0.577-0.623)	0.022 (0.000-0.623)	0.784 (0.780-0.789)	0.364 (0.359-0.369)	0.018 (0.017-0.020)	0.036 (0.034-0.038)
RETAIN	0.457 (0.431-0.483)	0.014 (0.000-0.483)	0.061 (0.059-0.064)	0.954 (0.952-0.956)	0.020 (0.018-0.021)	0.030 (0.028-0.032)
BEHRT (No pretraining)	0.554 (0.531-0.578)	0.017 (0.000-0.578)	0.584 (0.579-0.589)	0.504 (0.499-0.509)	0.018 (0.016-0.019)	0.034 (0.032-0.036)
BEHRT (Pretrained)	0.543 (0.518-0.567)	0.018 (0.000-0.567)	0.540 (0.535-0.545)	0.538 (0.533-0.543)	0.018 (0.016-0.019)	0.034 (0.032-0.036)
CLMBR-t-base (Pretrained)	0.545 (0.521-0.570)	0.018 (0.000-0.570)	0.494 (0.489-0.500)	0.587 (0.582-0.592)	0.018 (0.017-0.019)	0.035 (0.033-0.037)
CDM-BERT (No pretraining)	0.656 (0.630-0.682)	0.047 (0.021-0.074)	0.540 (0.535-0.545)	0.702 (0.697-0.706)	0.027 (0.025-0.029)	0.051 (0.049-0.054)
CDM-BERT (Pretrained)	0.893 (0.875-0.911)	0.257 (0.239-0.275)	0.796 (0.792-0.800)	0.916 (0.913-0.919)	0.126 (0.122-0.129)	0.217 (0.213-0.222)

*Abbreviations: SNUH=Seoul National University Hospital, SNUBH=Seoul National University Bundang Hospital, CNUH=Chungbuk National University Hospital, AUROC=area under the receiver operating characteristic curve, AUPRC=area under the precision-recall curve

*Bold indicates the best performance.

*The confidence intervals of the AUROC and AUPRC were calculated using DeLong's method, and those of sensitivity, specificity, precision, and F1-score were calculated using Wilson's method

Feature importance

Feature importance was assessed using fine-tuned CDM-BERT, which demonstrated the best prediction performance (Table 2). To evaluate whether the model focuses on clinically relevant features, we extracted attention scores from the model and compared averaged scores between CADR-related and unrelated concepts. CADR-related diagnosis

concepts included cancer, chronic kidney disease, and chronic liver disease, while CADR-unrelated diagnosis concepts included obesity, cataract, hypertension, and fracture. Similarly, CADR-related drug concepts included antibiotics, anticonvulsants, nonsteroidal anti-inflammatory drugs, and antitubercular drugs, while CADR-unrelated drug concepts included vitamins, normal saline, and dextrose. As shown in Figure 2a and 2b, attention scores of CADR-related concepts were higher than unrelated concepts for both diagnosis and drug domains across all institutions. Calculating attention scores for drug concepts of CNUH was not feasible because the drug concepts of CNUH were not compatible with those of SNUH. Figure 2c shows the 40 most important features of the model for each hospital. For SNUH, drugs (such as famotidine) and antibiotics (such as ceftriaxone and piperacillin) were the most important variables. For SNUBH, midazolam, famotidine, and fluid-related drugs and chest CT procedure items were the most important variables. For CNUH, measurements such as protein, albumin, and liver enzyme levels and conditions such as chronic kidney disease and acute kidney injury were the most important variables. No drugs appeared in the top 40 features for CNUH because it used drug codes different from those of SNUH. Accordingly, when predicting the CADRs in the CNUH cohort, the model could not utilize the drug records because they were replaced with the [UNK] code, which represents codes not found in the vocabulary. The number of codes and shared codes for each domain is summarized in Supplementary Table 4.

Sub-analysis of immediate and delayed CADR types

For a more detailed analysis, the CADRs were classified into immediate and delayed types. Furthermore, the prediction time point for the delayed type was divided into two categories: immediately before antibiotic treatment and immediately before CADR occurrence to assess the impact of cumulative antibiotic treatment history on CADR. Across all cohorts, the model demonstrated greater confidence in predicting the delayed type CADRs compared with the immediate type. The detailed performance metrics of the models are summarized in Figure 3 and Supplementary Table 5. The ROC and PR curves of the models are presented in Supplementary Figure 2. The feature importance results for each group are presented in Supplementary Figures 3–5.

For uncertainty analysis, we reported the results of conformal prediction analysis in Table 3. Considering the class imbalance of our cohort, we adopted Mondrian conformal prediction³³, which is class-conditional, with a significance level of $\alpha=0.05$. The average set size was smaller when the model predicted delayed-type CADR

compared to immediate-type CADR. For delayed-type CADR, the fraction of uncertain predictions was 0 in SNUH and SNUBH, and 0.144 in CNUH. The overall and negative coverages were acceptable and close to 0.95, while positive coverage showed slight variability (0.915-0.950). This phenomenon is consistent with previous findings that the minority class tends to be undercovered in imbalanced datasets.^{34,35}

Table 3. Conformal prediction analysis of immediate and delayed CADR types

	Coverage	Average set size	Fraction of uncertain predictions	Coverage (positive)	Coverage (negative)
Internal dataset (SNUH)					
Immediate type	0.951 $\pm$ 0.002	1.034 $\pm$ 0.026	0.034 $\pm$ 0.026	0.925 $\pm$ 0.022	0.951 $\pm$ 0.002
Delayed type (before Abx Tx)	0.952 $\pm$ 0.001	0.995 $\pm$ 0.003	0.000 $\pm$ 0.000	0.939 $\pm$ 0.038	0.952 $\pm$ 0.001
Delayed type (before CADR)	0.949 $\pm$ 0.003	0.983 $\pm$ 0.003	0.000 $\pm$ 0.000	0.937 $\pm$ 0.023	0.950 $\pm$ 0.003
External dataset 1 (SNUBH)					
Immediate type	0.949 $\pm$ 0.001	1.301 $\pm$ 0.021	0.301 $\pm$ 0.021	0.915 $\pm$ 0.024	0.950 $\pm$ 0.001
Delayed type (before Abx Tx)	0.951 $\pm$ 0.001	1.000 $\pm$ 0.001	0.000 $\pm$ 0.000	0.935 $\pm$ 0.017	0.951 $\pm$ 0.001
Delayed type (before CADR)	0.950 $\pm$ 0.003	1.000 $\pm$ 0.001	0.000 $\pm$ 0.000	0.935 $\pm$ 0.012	0.950 $\pm$ 0.003
External dataset 2 (CNUH)					
Immediate type	0.952 $\pm$ 0.002	1.766 $\pm$ 0.012	0.766 $\pm$ 0.012	0.950 $\pm$ 0.012	0.952 $\pm$ 0.002
Delayed type (before Abx Tx)	0.950 $\pm$ 0.001	1.299 $\pm$ 0.016	0.299 $\pm$ 0.016	0.924 $\pm$ 0.008	0.950 $\pm$ 0.001
Delayed type (before CADR)	0.950 $\pm$ 0.002	1.144 $\pm$ 0.073	0.144 $\pm$ 0.073	0.925 $\pm$ 0.026	0.950 $\pm$ 0.002

*Abbreviations: SNUH=Seoul National University Hospital, SNUBH=Seoul National University Bundang Hospital, CNUH=Chungbuk National University Hospital, Abx=antibiotics, Tx=treatment, CADR=cutaneous adverse drug reaction, AUROC=area under the receiver operating characteristic curve, AUPRC=area under the precision-recall curve

*Values are reported as mean $\pm$ standard deviation across five random seeds with a test size of 0.5.

Discussion

In this study, we developed, externally validated, and qualitatively analyzed CADR prediction models using patient trajectories and EHR FMs as well as other baselines. To extract high-quality patient records of skin rashes, we integrated nursing records with structured EHR data. CDM-BERT outperformed all other baselines, confirming the effectiveness of DE for predicting CADR. Our model showed remarkable generalizability, maintaining consistent performance across three different hospitals without requiring institution-specific finetuning. By analyzing CADR subtypes (immediate and delayed types), we found distinct predictive patterns that enhance applicability to clinical practice with subdivided use cases. These findings demonstrate the effectiveness of EHR FMs as a CADR prediction model that can directly contribute to patient care.

We assumed that the model pretrained using dense medical records from inpatients would reflect the variation in medical status before and after various treatments, including medication. Several studies have predicted ADRs using patient records from EHR data. One previous study developed a prediction model using a random forest algorithm, with clinical code data (prescription drugs and diagnoses), and measurements extracted from the EHR database.³⁶ Another study also aggregated diagnosis, drug, and measurement records to develop prediction models for various adverse drug events using several machine learning methods.³⁷ Nevertheless, only a few studies have used EHR FMs for ADR prediction.²³ In addition, to the best of our knowledge, CADR records have not been explored through EHR FMs.

Nursing records are important for developing an ADR prediction model, as they include detailed and regular observations of clinical signs and symptoms, such as skin rash, itching, nausea, and vomiting. Despite the abundance of information in nursing records, utilizing them from the EHR is challenging, as not all records are typically converted into diagnosis or observation codes. Many items in nursing records are ambiguous in matching existing diagnostic codes due to their variability in type, severity, and durability. Although some diagnostic records for skin conditions are present in our EHR databases, their integrity could not be guaranteed because patients in nursing records and EHR were hardly overlapped. In this study, we combined nursing records and EHR data to fully utilize the CADR records. Subsequently, we constructed reliable prediction models that captured changes in the skin condition of the patients before and after antibiotic treatment.

Many drugs were included in the top 40 most important features of SNUH and SNUBH; however, none were included in those of CNUH (Figure 2c), as CNUH uses different drug codes. Although all hospitals in the study operated observational medical outcomes partnership (OMOP) common data model (CDM)-based EHR databases and used the same terminologies for medical codes, the specific classes for drug codes were different. SNUH and SNUBH used drug codes in the "clinical drug" class of Rx-Norm, whereas CNUH used those in the "quant clinical drug" class of Rx-Norm. For example, the Rx-Norm concept ID "19027493" indicates "famotidine 10 MG/ML Injectable Solution" under the "Clinical Drug" class, whereas the Rx-Norm concept ID "35606552" indicates "2 ML famotidine 10 MG/ML Injection" under the "Quant Clinical Drug" class. These codes are almost identical but the model recognizes them as different. However, although drug information from the CNUH was not available for external validation, our model still achieved an AUROC of approximately 0.9. Instead of drug information, the model classified

patients with CADRs primarily based on the measurement items, which was hardly observed in SNUH and SNUBH. Additionally, we also conducted additional experiments using CDM-BERT after replacing drug codes with "unknown" codes, and the model achieved AUROCs of 0.971 and 0.928 for SNUH and SNUBH, respectively, demonstrating minimal performance degradation compared to the results of original setting (0.975 and 0.928). This implies that our model does not solely rely on a specific set of medical codes but comprehensively considers the overall patient information.

Our feature importance analysis revealed that our model assigns higher attention scores to clinically relevant concepts such as cancer, chronic kidney disease, and chronic liver disease compared to unrelated concepts (Figure 2a and 2b). This finding suggests that our model makes the decision focusing more on clinically relevant concepts. However, it is important to note that attention scores indicate how much the corresponding records influenced the decision of model and high attention scores do not directly indicate clinically meaningful records. In addition, as described in above paragraph, EHR FM makes decisions by comprehensively analyzing all given concepts, rather than relying solely on specific "meaningful" concepts or domains. This makes the interpretation of EHR FMs difficult despite their superior performance.

Although skin tests can be performed for patients with suspected β -lactam allergy, they are not routinely recommended before antibiotic administration in patients without a history of hypersensitivity. While standardized protocols exist for β -lactam allergy testing, the diagnostic sensitivity varies depending on factors such as the time elapsed since the reaction and the specific drug involved³⁸⁻⁴⁰. In a study evaluating the clinical utility of beta-lactam antibiotics skin testing, where skin tests were performed with five penicillin and cephalosporin antibiotics in patients with no history of allergy to beta-lactam antibiotics, 97 (6.8%) of the 1,421 patients tested positive for at least one drug. However, none of these patients showed hypersensitivity to cephalosporins. However, 4 of the 1,324 patients with negative test results exhibited hypersensitivity following cephalosporin injection, indicating that skin testing before the use of penicillin and cephalosporin may not be clinically useful⁴¹. A study evaluating the effectiveness of skin testing prior to cefazolin use in 13,153 cases also failed to prove the clinical effectiveness of skin tests. Of the 13,153 individuals who underwent skin tests, 184 (1.4%) exhibited positive results. Of these, 174 avoided cefazolin use, and 10 underwent a stepwise challenge test with cefazolin, all of which were negative⁴². A meta-analysis of penicillin skin tests reported 30.7% sensitivity and 96.8% specificity of the skin tests, with an AUROC value of 0.686⁴³.

Delayed-type CADR are a broad spectrum of diseases that occur after a considerable period of time after antibiotic exposure and have various pathophysiological and clinical risk factors. These delayed-type reactions can sometimes lead to severe and potentially life-threatening conditions. Specifically, systemic manifestations occur in 30% of CADR cases, with serious reactions such as SJS, TEN, and general exfoliative dermatitis in 5.2% of cases⁴. Furthermore, delayed CADR are clinically difficult to predict, because there is no method for predicting delayed reactions through testing, like immediate reactions. For this problem, our findings showed that EHR FMs with DE demonstrated superior and more stable prediction performance for delayed reactions over the immediate reactions. This suggests that baseline patient information holds greater predictive value for delayed-type reactions and may include generalized risk factors relevant across hospitals. In addition, the performance gap between delayed and immediate CADR is noteworthy, as we did not provide separate information on immediate and delayed types. Our model itself recognized the disparity between different CADR types and demonstrated validity across all three cohorts. This finding indicates that our model can derive meaningful insights based on patient medical history, even without explicit information provided.

Among delayed-type CADR, severe CADR such as DRESS syndrome, SJS, and TEN are rare but potentially life-threatening. The incidence of SJS and TEN is estimated to be 2 per 1 million people, whereas the incidence of DRESS syndrome is estimated to be 1 per 1000 to 1 per 10,000 people.⁴⁴ The diagnosis of severe CADR depends on the clinical presentation of the rash, its duration, associated symptoms (e.g., fever, pruritus, lymphadenopathy), and the time from drug administration to symptom onset. Direct immunofluorescence testing of the blistering rash and skin biopsy are necessary to rule out other possible causes. Clinically, it would be very useful to be able to identify and predict those at risk of progressing to severe ADRs among severe delayed CADR. However, the training data for our AI model is insufficient to distinguish between the severity and phenotype of skin rashes, making it impossible to identify and predict severe ADRs. In this study, we are the first to conduct research using AI to predict antibiotic-associated CADR, and we may also conduct future research to develop an AI model that can distinguish and predict severe CADR.

This study has some limitations. First, the code disparity between hospitals was not fundamentally addressed. Instead of matching different codes, we simply replaced the codes out of vocabulary with a [MASK] token. Although this approach did not significantly reduce the performance, a more robust strategy to handle code disparity is

imperative to enhance the compatibility of the model. Second, while our feature importance analysis demonstrated that our model focuses more on clinically relevant features, further interpretation of the complex attention patterns is needed to validate the utility of attention score analysis in EHR FMs. Third, CADR records are usually clinically classified by more complicated conditions than our settings. CADR records are classified into immediate and delayed reactions based on the 1-6h window after drug administration^{45,46}. Immediate reactions manifest as urticaria and angioedema, whereas delayed reactions mainly manifest as maculopapular rashes. A more detailed analysis of these subtypes and the corresponding application strategies is suggested for future studies. Fourth, this study involved hospitals from a single country, and external validation using a global cohort is necessary for model generalization. Fifth, we extracted the CADR records using nursing reports and rule-based regular expressions. This approach may have missed some records and may not be applicable to other institutions, as nursing reports are not strictly structured and format varies among institutions. Therefore, leveraging large LMs could enhance the CADR record extraction process. Sixth, this study focused on mild CADR records, including skin lesions, rashes, urticaria, and eruptions. A more comprehensive study design that covers diverse CADR records is required. Given the demonstrated effectiveness of EHR FMs for predicting mild CADR records, predicting severe CADR records, such as SJS and TEN, and identifying new factors causing those CADR records are suggested for future work. Seventh, our study compared patients who developed rashes during antibiotic treatment to those who did not, rather than comparing them with patients who developed rashes without antibiotic exposure. Future studies comparing "antibiotic with rash" versus "rash without antibiotic exposure" could help distinguish antibiotics-associated reactions from general rash susceptibility. Eighth, our current model predicts overall CADR risk rather than identifying specific culprit antibiotics or drug classes. Since drug allergy typically develops to specific drugs or drug classes rather than overall antibiotics in general, future studies need to develop drug class-specific prediction models to provide more valuable clinical guidance for antibiotic selection. Nevertheless, the current model can serve as a risk stratification tool to identify patients who may benefit from enhanced monitoring during antibiotic treatment.

Methods

Data curation

This study used data from SNUH between December 2012 and April 2023, SNUBH between April 2003 and

December 2022, and CNUH between July 2020 and September 2024. All included hospitals operate an EHR database based on the OMOP CDM⁴⁷. We collected data from four domains (tables) of the CDM schema: condition, drug, measurement, and procedure. Terminologies for diagnosis, procedure, measurement, and prescription were based on SNOMED CT⁴⁸, Rx-Norm and Rx-Norm Extension⁴⁹, and LOINC⁵⁰, respectively. In addition to EHR data, we combined nursing statements and reports to process information regarding skin rashes.

For the internal dataset (SNUH), we randomly split the data into training (70%), validation (10%), and hold-out test (20%) datasets. Training and validation datasets were used to develop the models, including LM pretraining. External datasets were used to validate the trained models. As data sharing across hospitals is strictly prohibited, we only exported the weights of the trained models, and external validation was performed in each hospital.

The institutional review boards of the SNUH (approval no. E-2403-088-1521), SNUBH (approval no. X-2403-890-901), and CNUH (approval no. 2024-06-022-001) approved the study and waived the requirement for informed consent, as only retrospective and observational EHR data was used. The study adhered to the principles outlined in the Declaration of Helsinki, the Korean Bioethics and Safety Act (Law No. 16372), and the Human Research Protection Program–Standard Operating Procedure of the Seoul National University Hospital.

Nursing statements and reports

To obtain high-quality skin rash records, we used nursing statements and reports from each hospital. Note that "nursing statements" and "nursing reports" are separate terms in this study. The nursing records of SNUH and SNUBH are based on nursing statements, which are based on a code system. With this system, nurses directly enter the codes based on the status or condition of patients, instead of writing free-text nursing reports. The examples of the nursing statements are "rash is present," "urticaria is present," "no rash is observed," and "skin lesion is present," and each statement has an assigned code. We extracted all statements including the keywords "skin lesion," "rash," "urticaria," and "eruption" and excluded inappropriate nursing statements including "diaper rash," "abrasion," and "lacerated," as these conditions are not likely to be associated with antibiotic-associated CADR. Five allergists confirmed that included nursing statements like "rash is present," "urticaria is present," and "skin lesion is present" were appropriate.

In contrast, CNUH does not have a structured nursing statements system and only has natural language-

based nursing reports. As the nursing reports were written in plain Korean text and were less structured, we developed rule-based regular expression algorithms. The detailed rules are summarized in Supplementary Table 6. We randomly selected more than 1,000 samples that were classified as having skin rash or not, and five allergists confirmed that all sample reports were accurate.

Prediction timepoint

To train the CADR prediction model in a practical scenario, we considered the most recent inpatient visit where antibiotic treatment was provided, and the prediction time point was set as the antibiotic prescription immediately preceding the CADR occurrence. The reasons for the prediction timepoint definition are as follows: First, we considered predicting the first CADR for each patient, but the prediction performance deteriorated because of fewer records. Second, we attempted to rule out potential bias due to the conditions for the prediction timepoint. Accordingly, we designed the model to adapt to various situations, whether patient history records are abundant or not. We also considered three additional scenarios to validate the CADR prediction model: predicting the immediate and delayed CADR at the initiation of the antibiotic treatment and predicting the delayed CADR just before the CADR occurrence. The CADR prediction process and timepoints are presented in Figure 4.

Model training

To train the tree models as baselines, we transformed the patient trajectories into count-based one-dimensional vectors, referring to previous studies. Time binning counts the occurrences of a code in different time intervals. We used intervals of 0–1 day, 1–7 days, and 7+ days from the prediction timepoint. The count-based vector was then concatenated with age and sex information. Hyperparameter tuning was performed using grid search (Supplementary Table 7). Medical codes that were not in the SNUH vocabulary were omitted from the external validation. The tree models were trained and validated using the Scikit-learn package (version 1.0.2) in Python (version 3.8.10).

In addition to tree-based machine learning models, we used several deep learning models, including RNN-based models (LSTM and GRU), modified RNN-based networks with attention mechanisms (RETAIN), and EHR FMs (BEHRT, CLMBR-t-base, and CDM-BERT). BEHRT is a basic BERT-based EHR foundation model and CLMBR-t-base is an open weight EHR foundation model that generates patient representations which are

subsequently finetuned by LightGBM⁵¹. CDM-BERT is a modified BEHRT with DE for the OMOP CDM format. All EHR FMs were based on pretraining-finetuning framework. All deep learning models were based on the same embedding of patient records and shared the same embedding layers but different subsequent layers. The embedding sizes of the token, age, segment, and domain were set to 50000, 180, 1024, and 20, respectively. All deep learning models used six RNN or self-attention layers with a hidden dimension of 256. The BERT-based LM used eight attention heads. To train the CADR prediction models, we added a feed-forward neural network (FFNN), a hyperbolic tangent (Tanh) activation function, dropout, and another FFNN to the representation of the first token to yield a vector of size two, and binary cross entropy was used to minimize the loss function. We calculated the AUROC values for every 1000 batches, and early stopping was implemented if the AUROC did not improve for 20 cycles. To address class imbalance, we randomly oversampled the patients with CADRs to one-tenth of those without CADRs. The batch size, learning rate, and dropout rate were 16, 5e-5, and 0.1, respectively. We used the pretrained models from our previous study and confirmed that the hold-out test dataset was entirely independent of the pretraining dataset. The hyperparameter tuning for LightGBM in CLMBR-t-base was performed using the same grid search approach as described for random forest and GBM (Supplementary Table 7). For the deep learning and BERT implementation, we used PyTorch (version 1.12.0) and the HuggingFace package (version 4.41.2)⁵² in Python (version 3.8.10). The code implementation for the model training was based on the GitHub repository (<https://github.com/kicarussays/CDM-BERT>).

Feature importance

We conducted two feature importance analyses: (1) comparing attention scores between CADR-related and unrelated diagnosis concepts, and (2) identifying the top 40 most important features. For the attention score comparison, we extracted and averaged attention scores for predefined concept groups. For identifying important features, the calculating process involved four steps: First, we averaged the attention matrices of eight attention heads from the last self-attention layer of the LM, yielding a simplified 2048 by 2048 attention matrix A . For each attention matrix A , an attention score A_{ij} indicates how closely i^{th} and j^{th} tokens are related.⁵³ Note that $\sum_j A_{ij} = 1 \forall i \in (0, \dots, 2048)$. Second, we averaged the simplified attention matrix across rows, resulting in a one-dimensional vector of length 2048, representing the importance of each token. Third, for each patient, we identified the top 10% of tokens

with the highest average attention scores. Finally, for each variable, we counted the number of patients who included that variable in the top 40 features. We used count-based feature importance to improve interpretability, instead of the trajectory attention vector proposed in our previous study.

Statistical analysis

Characteristics (age, sex, and comorbidities) between the CADR and non-CADR groups in all cohorts were compared by calculating P-values using Student's t-test for age and Fisher's exact test for other variables. We reported the AUROC, AUPRC, sensitivity, specificity, precision, and F1-score as performance metrics. The confidence intervals of AUROC and AUPRC were calculated using the DeLong's method⁵⁴, whereas those of the F1-score, sensitivity, specificity, and precision were calculated using the Wilson's method⁵⁵. Statistical significance was set at $\alpha = 0.05$. All statistical analyses were performed using Scikit-learn (version 1.0.2) in Python (version 3.8.10).

References

- 1 Edwards, I. R. & Aronson, J. K. Adverse drug reactions: definitions, diagnosis, and management. *Lancet* **356**, 1255–1259 (2000). [https://doi.org/10.1016/s0140-6736\(00\)02799-9](https://doi.org/10.1016/s0140-6736(00)02799-9)
- 2 Lazarou, J., Pomeranz, B. H. & Corey, P. N. Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. *Jama* **279**, 1200–1205 (1998). <https://doi.org/10.1001/jama.279.15.1200>
- 3 Seo, B. *et al.* Incidence and Economic Burden of Adverse Drug Reactions in Hospitalization: A Prospective Study in Korea. *J Korean Med Sci* **38**, e56 (2023). <https://doi.org/10.3346/jkms.2023.38.e56>
- 4 Gomes, E. R. & Demoly, P. Epidemiology of hypersensitivity drug reactions. *Curr Opin Allergy Clin Immunol* **5**, 309–316 (2005). <https://doi.org/10.1097/01.all.0000173785.81024.33>
- 5 Thong, B. Y., Leong, K. P., Tang, C. Y. & Chng, H. H. Drug allergy in a general hospital: Results of a novel prospective inpatient reporting system. *Ann Allergy Asthma Immunol* **90**, 342–347 (2003). [https://doi.org/10.1016/s1081-1206\(10\)61804-2](https://doi.org/10.1016/s1081-1206(10)61804-2)
- 6 Frey, N. *et al.* The Epidemiology of Stevens-Johnson Syndrome and Toxic Epidermal Necrolysis in the UK. *J Invest Dermatol* **137**, 1240–1247 (2017). <https://doi.org/10.1016/j.jid.2017.01.031>
- 7 Lee, E. Y., Knox, C. & Phillips, E. J. Worldwide Prevalence of Antibiotic-Associated Stevens-Johnson Syndrome and Toxic Epidermal Necrolysis: A Systematic Review and Meta-analysis. *JAMA Dermatol* **159**, 384–392 (2023). <https://doi.org/10.1001/jamadermatol.2022.6378>
- 8 Singh, M. P. *et al.* "Antimicrobial utilization in a paediatric intensive care unit in India: A step towards strengthening antimicrobial stewardship practices". *PLOS ONE* **19**, e0310515 (2024). <https://doi.org/10.1371/journal.pone.0310515>
- 9 Rossi, G., da Silva Cartell, A. & Marchiori Bakos, R. Dermoscopic Aspects of Cutaneous Adverse Drug Reactions. *Dermatology Practical & Conceptual* **11**, e2021136 (2021). <https://doi.org/10.5826/dpc.1101a136>
- 10 Finkelstein, Y., Macdonald, E. M., Li, P., Hutson, J. R. & Juurlink, D. N. Recurrence and mortality following severe cutaneous adverse reactions. *Jama* **311**, 2231–2232 (2014). <https://doi.org/10.1001/jama.2014.839>
- 11 Duong, T. A., Valeyrie-Allanore, L., Wolkenstein, P. & Chosidow, O. Severe cutaneous adverse reactions to drugs. *Lancet* **390**, 1996–2011 (2017). [https://doi.org/10.1016/s0140-6736\(16\)30378-6](https://doi.org/10.1016/s0140-6736(16)30378-6)
- 12 Parasrampur, S. & Henry, J. Hospitals' use of electronic health records data, 2015–2017. *ONC Data Brief* **46**, 13 (2019).
- 13 Liang, J. *et al.* Adoption of electronic health records (EHRs) in China during the past 10 years: consecutive survey data analysis and comparison of sino-american challenges and experiences.

- Journal of medical Internet research* **23**, e24813 (2021). <https://doi.org/10.2196/24813>
- 14 Lee, K. *et al.* Digital Health Profile of South Korea: A Cross Sectional Study. *International Journal of Environmental Research and Public Health* **19**, 6329 (2022).
 - 15 Huang, D., Cogill, S., Hsia, R. Y., Yang, S. & Kim, D. Development and external validation of a pretrained deep learning model for the prediction of non-accidental trauma. *npj Digital Medicine* **6**, 131 (2023). <https://doi.org/10.1038/s41746-023-00875-y>
 - 16 Shang, J., Ma, T., Xiao, C. & Sun, J. Pre-training of graph augmented transformers for medication recommendation. *arXiv preprint arXiv:1906.00346* (2019). <https://doi.org/https://doi.org/10.48550/arXiv.1906.00346>
 - 17 Li, Y. *et al.* BEHRT: Transformer for Electronic Health Records. *Scientific Reports* **10**, 7155 (2020). <https://doi.org/10.1038/s41598-020-62922-y>
 - 18 Rasmy, L., Xiang, Y., Xie, Z., Tao, C. & Zhi, D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digital Medicine* **4**, 86 (2021). <https://doi.org/10.1038/s41746-021-00455-y>
 - 19 Pang, C. *et al.* in *Proceedings of Machine Learning for Health* Vol. 158 (eds Roy Subhrajit *et al.*) 239—260 (PMLR, Proceedings of Machine Learning Research, 2021).
 - 20 Li, Y. *et al.* Hi-BEHRT: Hierarchical Transformer-Based Model for Accurate Prediction of Clinical Events Using Multimodal Longitudinal Electronic Health Records. *IEEE Journal of Biomedical and Health Informatics* **27**, 1106–1117 (2023). <https://doi.org/10.1109/JBHI.2022.3224727>
 - 21 Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018). <https://doi.org/https://doi.org/10.48550/arXiv.1810.04805>
 - 22 Wornow, M., Thapa, R., Steinberg, E., Fries, J. & Shah, N. Ehrshot: An ehr benchmark for few-shot evaluation of foundation models. *Advances in Neural Information Processing Systems* **36**, 67125–67137 (2023). <https://doi.org/https://doi.org/10.48550/arXiv.2307.02028>
 - 23 Kim, J. *et al.* Pretrained patient trajectories for adverse drug event prediction using common data model-based electronic health records. *Communications Medicine* **5**, 232 (2025). <https://doi.org/10.1038/s43856-025-00914-7>
 - 24 Fiszenson-Albala, F. *et al.* A 6-month prospective survey of cutaneous drug reactions in a hospital setting. *Br J Dermatol* **149**, 1018–1022 (2003). <https://doi.org/10.1111/j.1365-2133.2003.05584.x>
 - 25 Hernández-Salazar, A. *et al.* Epidemiology of adverse cutaneous drug reactions. A prospective study in hospitalized patients. *Arch Med Res* **37**, 899–902 (2006). <https://doi.org/10.1016/j.arcmed.2006.03.010>
 - 26 Park, C. S. *et al.* The use of an electronic medical record system for mandatory reporting of drug hypersensitivity reactions has been shown to improve the management of patients in the

- university hospital in Korea. *Pharmacoepidemiol Drug Saf* **17**, 919–925 (2008). <https://doi.org/10.1002/pds.1612>
- 27 Breiman, L. Random Forests. *Machine Learning* **45**, 5–32 (2001). <https://doi.org/10.1023/A:1010933404324>
- 28 Jerome, H. F. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* **29**, 1189–1232 (2001). <https://doi.org/10.1214/aos/1013203451>
- 29 Graves, A. in *Supervised Sequence Labelling with Recurrent Neural Networks* (ed Alex Graves) 37–45 (Springer Berlin Heidelberg, 2012).
- 30 Chung, J., Gulcehre, C., Cho, K. & Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).
- 31 Choi, E. *et al.* Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in neural information processing systems* **29** (2016). <https://doi.org/10.48550/arXiv.1608.05745>
- 32 Guo, L. L. *et al.* EHR foundation models improve robustness in the presence of temporal distribution shift. *Scientific Reports* **13**, 3767 (2023). <https://doi.org/10.1038/s41598-023-30820-8>
- 33 Vovk, V. Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence* **74**, 9–28 (2015). <https://doi.org/10.1007/s10472-013-9368-4>
- 34 Sun, J. *et al.* Applying Mondrian Cross-Conformal Prediction To Estimate Prediction Confidence on Large Imbalanced Bioactivity Data Sets. *Journal of Chemical Information and Modeling* **57**, 1591–1598 (2017). <https://doi.org/10.1021/acs.jcim.7b00159>
- 35 Ding, T., Angelopoulos, A., Bates, S., Jordan, M. & Tibshirani, R. J. Class-conditional conformal prediction with many classes. *Advances in neural information processing systems* **36**, 64555–64576 (2023).
- 36 Zhao, J., Henriksson, A., Asker, L. & Boström, H. Predictive modeling of structured electronic health records for adverse drug event detection. *BMC Med Inform Decis Mak* **15 Suppl 4**, S1 (2015). <https://doi.org/10.1186/1472-6947-15-s4-s1>
- 37 Bampa, M. & Papapetrou, P. Aggregate-Eliminate-Predict: Detecting Adverse Drug Events from Heterogeneous Electronic Health Records. *ArXiv* **abs/1907.06058** (2019).
- 38 Khan, D. A. *et al.* Drug allergy: A 2022 practice parameter update. *J Allergy Clin Immunol* **150**, 1333–1393 (2022). <https://doi.org/10.1016/j.jaci.2022.08.028>
- 39 Broyles, A. D. *et al.* Practical Guidance for the Evaluation and Management of Drug Hypersensitivity: Specific Drugs. *J Allergy Clin Immunol Pract* **8**, S16–s116 (2020). <https://doi.org/10.1016/j.jaip.2020.08.006>
- 40 Doña, I. *et al.* An algorithm for the diagnosis of beta-lactam allergy, 2024 update. *Allergy* **80**, 633–637 (2025). <https://doi.org/10.1111/all.16348>

- 41 Yoon, S. Y. *et al.* Validation of the cephalosporin intradermal skin test for predicting immediate hypersensitivity: a prospective study with drug challenge. *Allergy* **68**, 938–944 (2013). <https://doi.org/10.1111/all.12182>
- 42 Kwon, J. W. *et al.* Results of Intradermal Skin Testing with Cefazolin according to a History of Hypersensitivity to Antibiotics. *J Korean Med Sci* **34**, e319 (2019). <https://doi.org/10.3346/jkms.2019.34.e319>
- 43 Sousa-Pinto, B. *et al.* Accuracy of penicillin allergy diagnostic tests: A systematic review and meta-analysis. *J Allergy Clin Immunol* **147**, 296–308 (2021). <https://doi.org/10.1016/j.jaci.2020.04.058>
- 44 Kardaun, S. H. *et al.* Drug reaction with eosinophilia and systemic symptoms (DRESS): an original multisystem adverse drug reaction. Results from the prospective RegiSCAR study. *Br J Dermatol* **169**, 1071–1080 (2013). <https://doi.org/10.1111/bjd.12501>
- 45 Broyles, A. D., Banerji, A. & Castells, M. Practical Guidance for the Evaluation and Management of Drug Hypersensitivity: General Concepts. *J Allergy Clin Immunol Pract* **8**, S3–s15 (2020). <https://doi.org/10.1016/j.jaip.2020.08.002>
- 46 Brockow, K. *et al.* EAACI position paper on how to classify cutaneous manifestations of drug hypersensitivity. *Allergy* **74**, 14–27 (2019). <https://doi.org/10.1111/all.13562>
- 47 Biedermann, P. *et al.* Standardizing registry data to the OMOP Common Data Model: experience from three pulmonary hypertension databases. *BMC Medical Research Methodology* **21**, 238 (2021). <https://doi.org/10.1186/s12874-021-01434-3>
- 48 Benson, T. *Principles of health interoperability HL7 and SNOMED*. (Springer Science & Business Media, 2012).
- 49 Liu, S., Wei, M., Moore, R., Ganesan, V. & Nelson, S. RxNorm: prescription for electronic drug information exchange. *IT Professional* **7**, 17–23 (2005). <https://doi.org/10.1109/MITP.2005.122>
- 50 McDonald, C. J. *et al.* LOINC, a Universal Standard for Identifying Laboratory Observations: A 5-Year Update. *Clinical Chemistry* **49**, 624–633 (2003). <https://doi.org/10.1373/49.4.624>
- 51 Ke, G. *et al.* in *Proceedings of the 31st International Conference on Neural Information Processing Systems* 3149–3157 (Curran Associates Inc., Long Beach, California, USA, 2017).
- 52 Wolf, T. *et al.* Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* (2019). <https://doi.org/10.48550/arXiv.1910.03771>
- 53 Vaswani, A. *et al.* Attention is all you need. *Advances in neural information processing systems* **30** (2017).
- 54 DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics* **44**, 837–845 (1988). <https://doi.org/10.2307/2531595>
- 55 Wilson, E. B. Probable Inference, the Law of Succession, and Statistical Inference. *Journal of the*

American Statistical Association **22**, 209–212 (1927).
<https://doi.org/10.1080/01621459.1927.10502953>

ARTICLE IN PRESS

Figure 1. Participant enrollment flowchart. An internal dataset from Seoul National University Hospital (SNUH) was used for training and internal validation of the model. External datasets from Seoul National University Bundang Hospital (SNUBH) and Chungbuk National University Hospital (CNUH) were used for external validation.

Figure 2. Feature importance analysis. (a) Comparison of attention scores between CADR-related and unrelated diagnosis concepts. (b) Comparison of attention scores between CADR-related and unrelated drug concepts. Drug concept analysis for CNUH was not possible due to different RxNorm concept classes used (see Discussion). (c) Top 40 most important features for predicting CADRs. The bar color indicates the domain of the medical code (condition, drug, measurement, and procedure). SNUH, Seoul National University Hospital; SNUBH, Seoul National University Bundang Hospital; CNUH, Chungbuk National University Hospital.

Figure 3. Risk comparison of patients without CADR and with immediate or delayed CADRs. The green line in each boxplot indicates the median, and the length of each box indicates the interquartile range (IQR) of the logits for each group. The ends of the box correspond to the 25th (Q1) and 75th (Q3) percentiles, and the lines extend to $Q1 - 1.5 \text{ IQR}$ and $Q3 + 1.5 \text{ IQR}$. SNUH, Seoul National University Hospital; SNUBH, Seoul National University Bundang Hospital; CNUH, Chungbuk National University Hospital; CADR, cutaneous adverse drug reaction; Abx, antibiotics; Tx, treatment.

Figure 4. CADR prediction process and prediction timepoints. (a) CDM-based EHR data and nursing records (nursing reports and statements) were merged and used to construct patient trajectories. Patient records until the last antibiotics prescription before CADR occurrence were used to predict CADR within 1 day. (b) Prediction timepoint for immediate CADRs. (c) Prediction timepoints for delayed CADRs. Prediction timepoints 1 and 2 involved the use of the records until the first antibiotic prescription and the last antibiotic prescription before CADR occurrence.

Author contributions

SHK and JK contributed to the conceptualization and design of the study. JK handled and mainly analyzed the research data, and all authors interpreted the results. JK constructed the machine learning and deep learning models and conducted statistical analysis. JK and KK wrote the original draft of the paper. SHK and KK reviewed the clinical evidence of the study. KSK, SY, and MGK provided the data, and JK verified the quality of the data. JK, SHK, KSK, MGK and SY had full access to the raw data. All authors had the final responsibility to submit for publication.

Acknowledgements

This work was supported by the Korea Institute of Drug Safety & Risk Management and the AI Institute at Seoul National University. SHK receives funding from Korea Institute of Drug Safety & Risk Management and JK is supported by the fellowship program of the AI Institute at Seoul National University. The funder played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

Competing interests

All authors declare no competing interests.

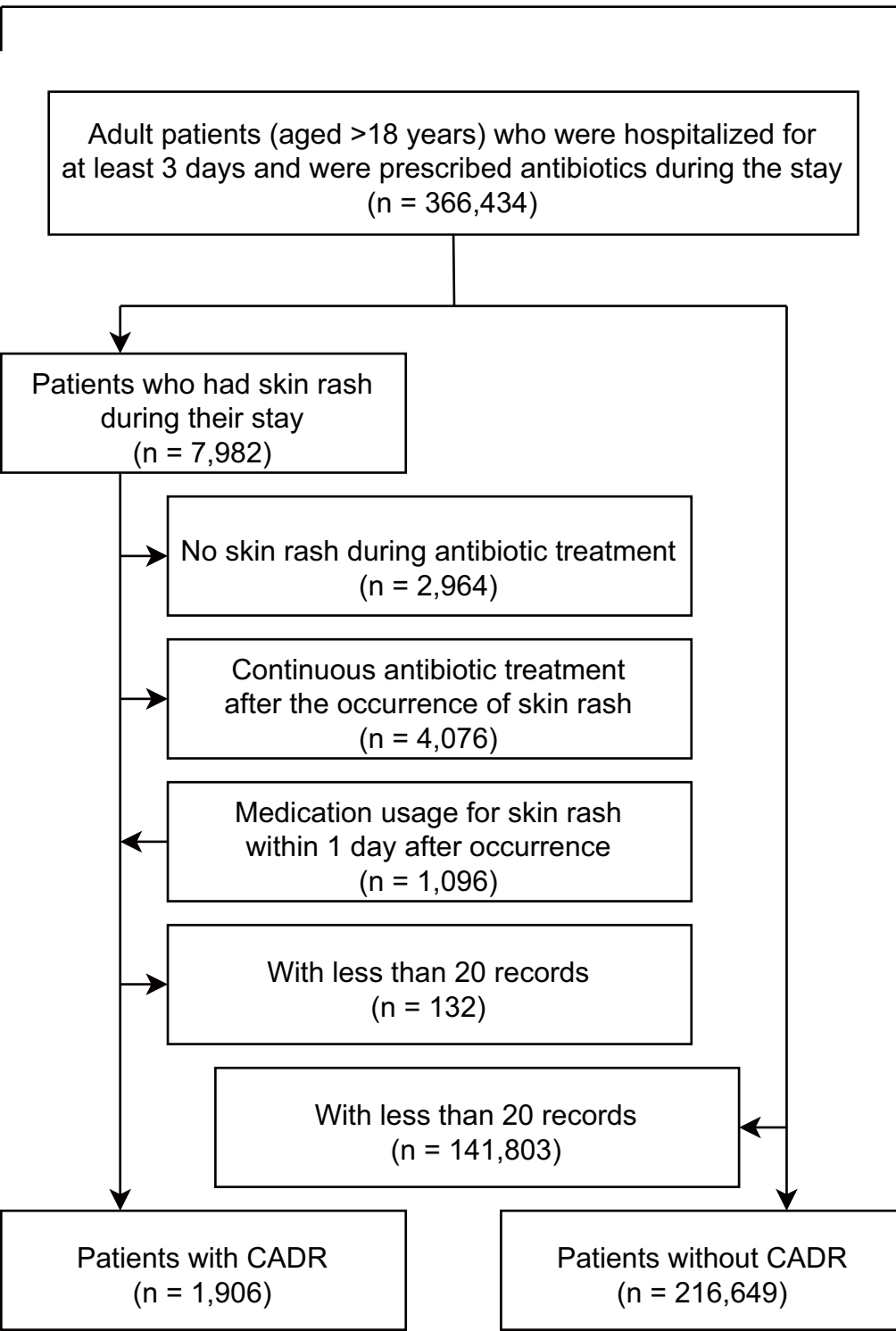
Data Availability

The raw data used in this study are not publicly available to preserve participant privacy. The data generated and analyzed during the study are available from the corresponding author upon reasonable request.

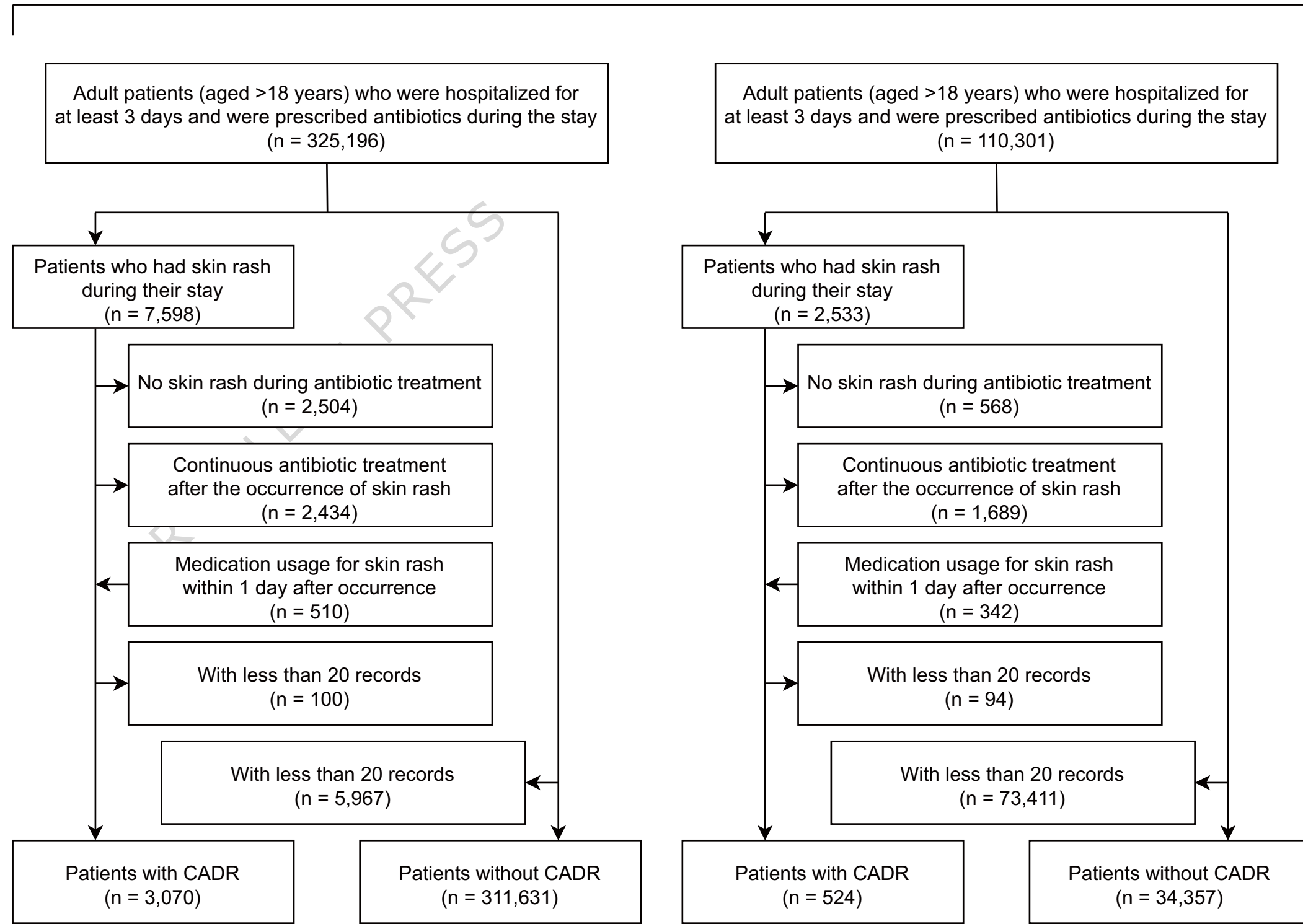
Code Availability

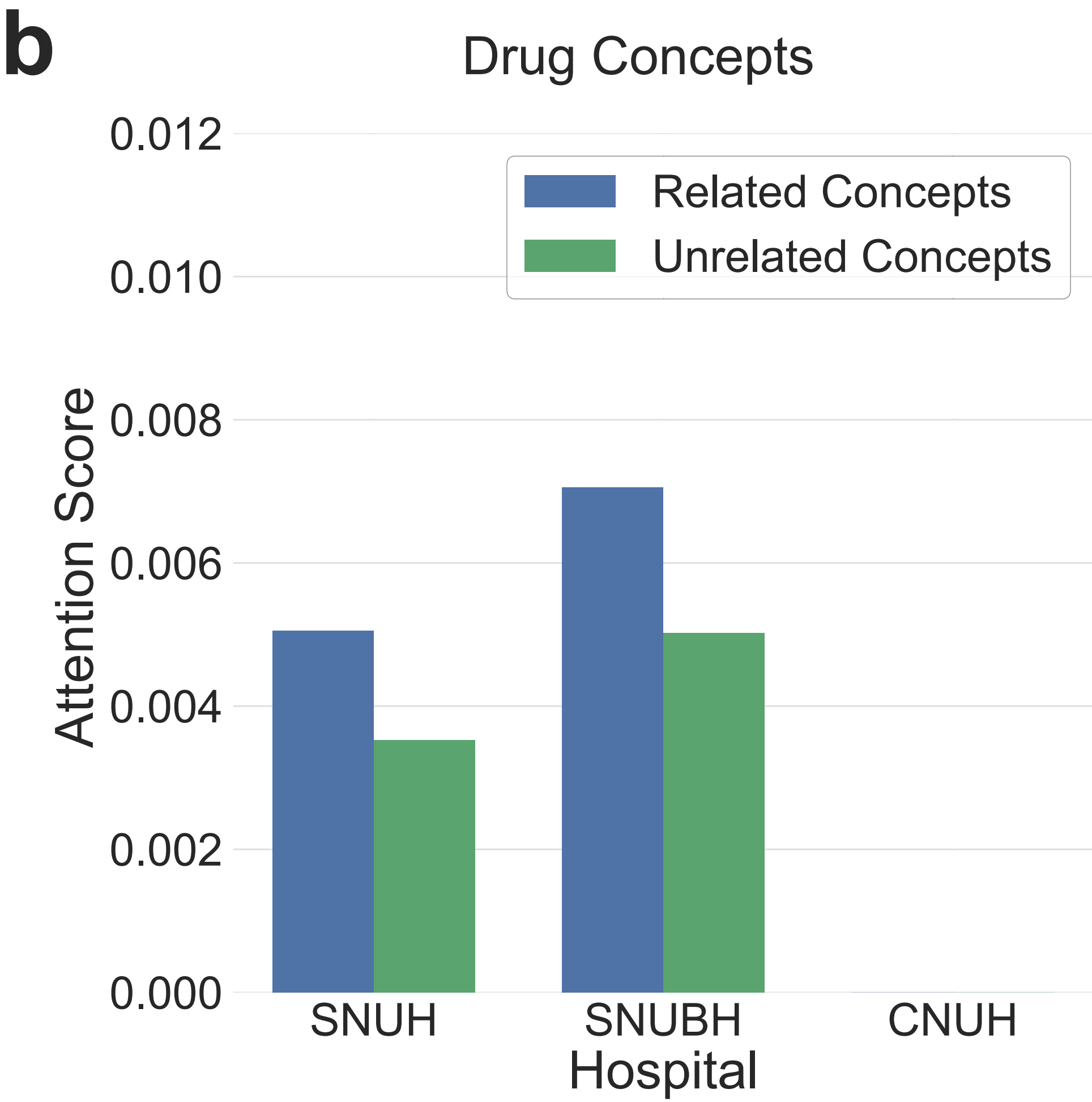
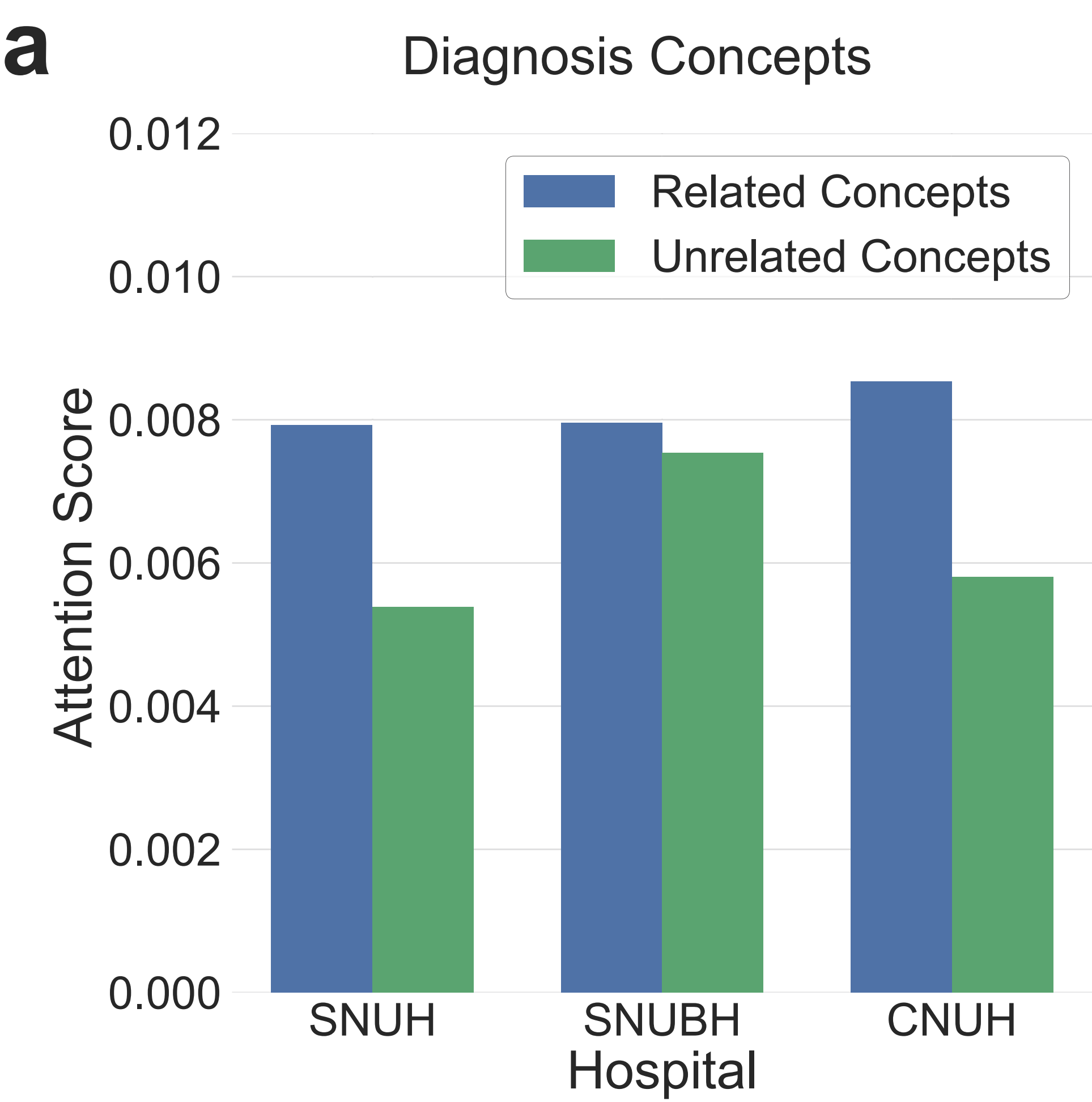
Source code for the experiments is publicly available at <https://github.com/kicarussays/CDM-BERT>.

Internal dataset (SNUH)



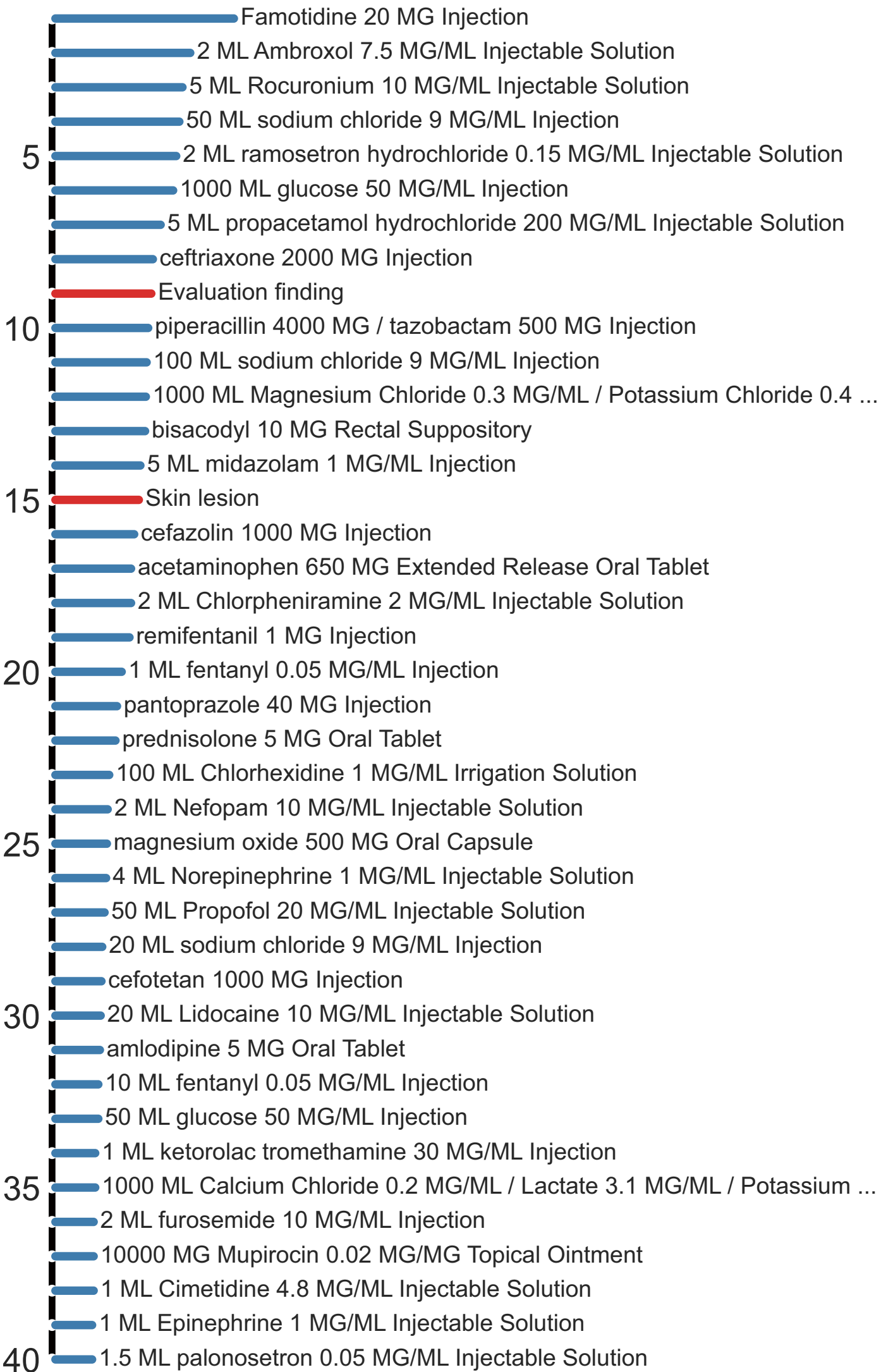
External datasets (SNUBH and CNUH)



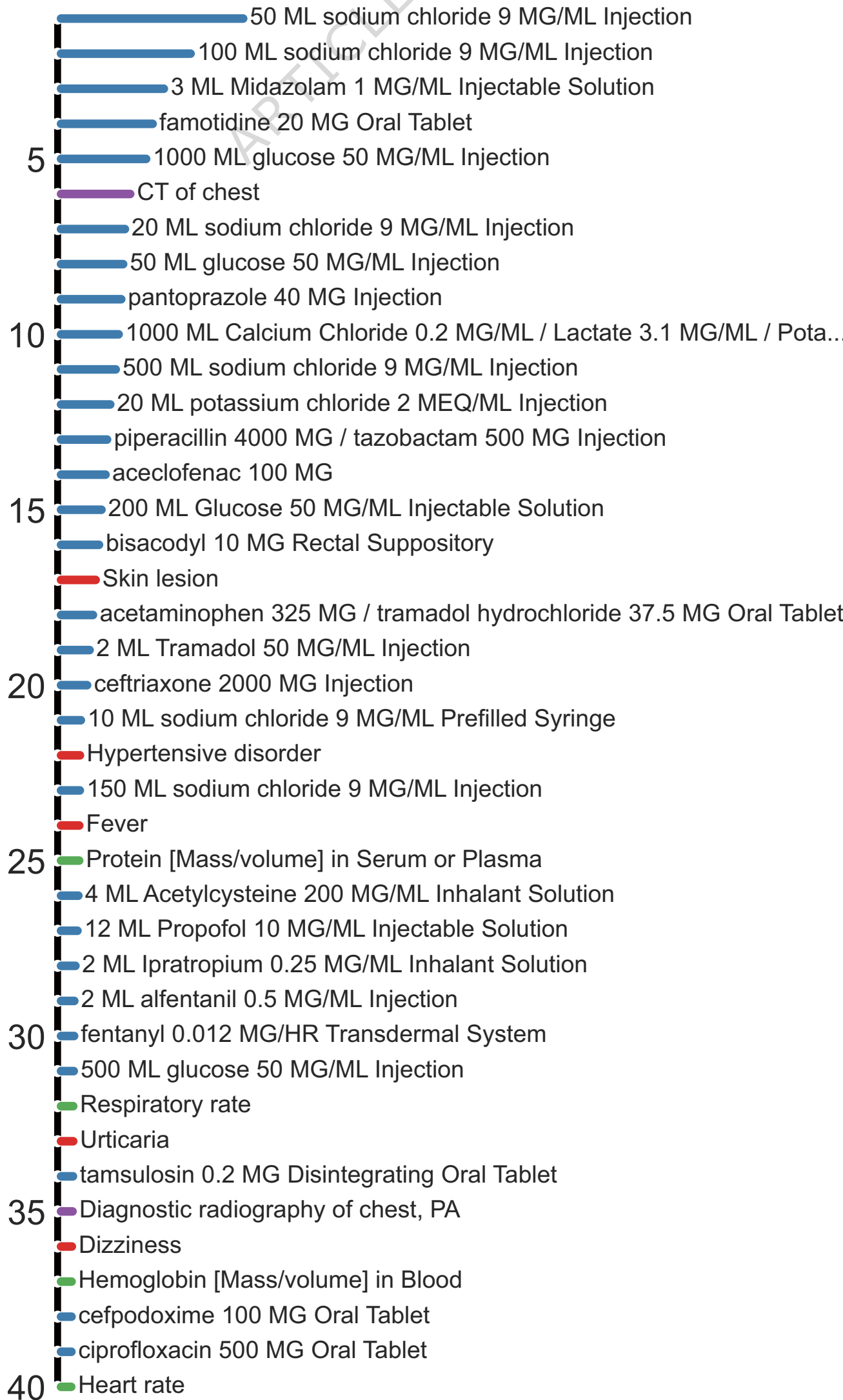


c

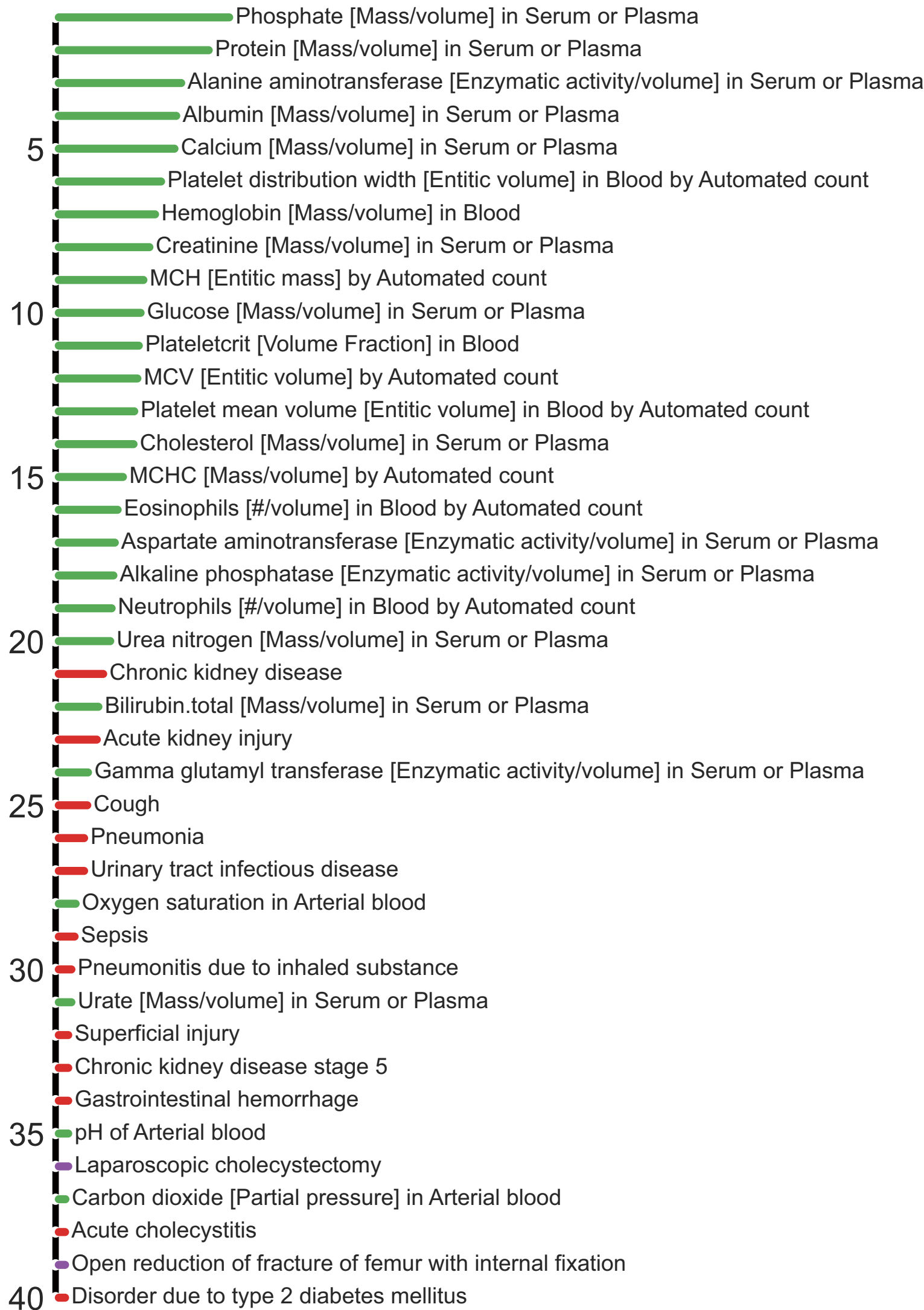
SNUH



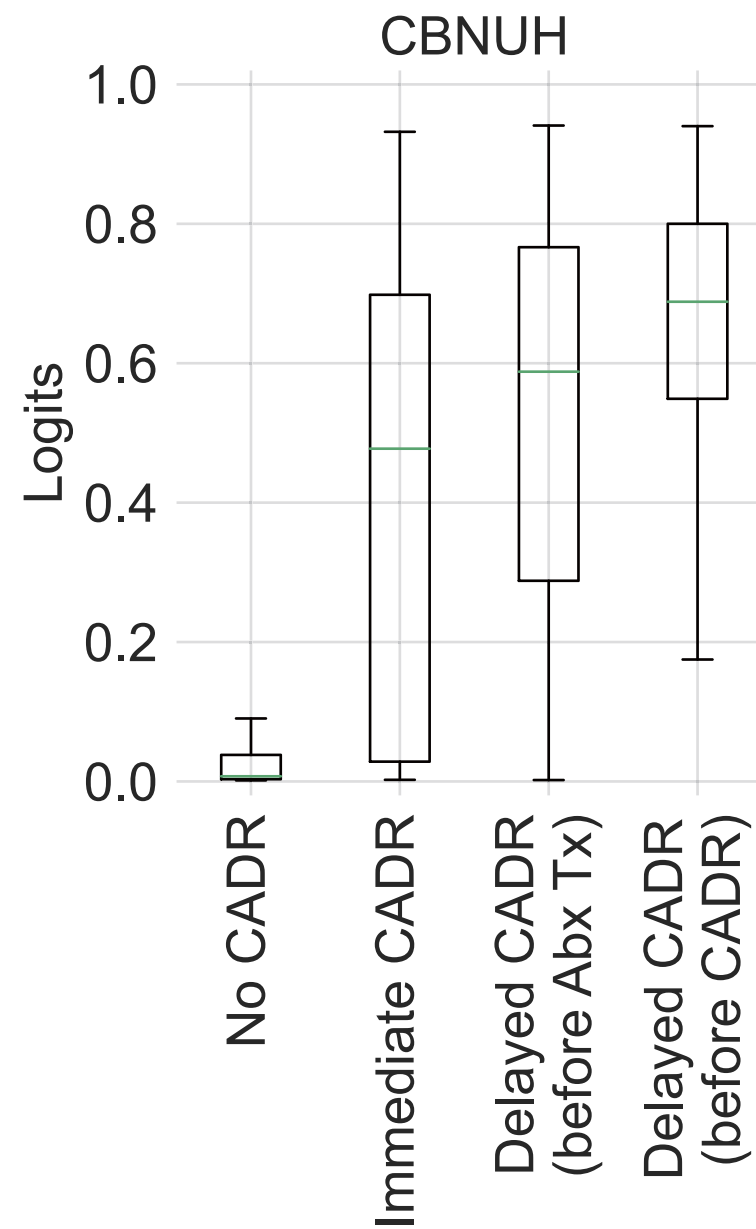
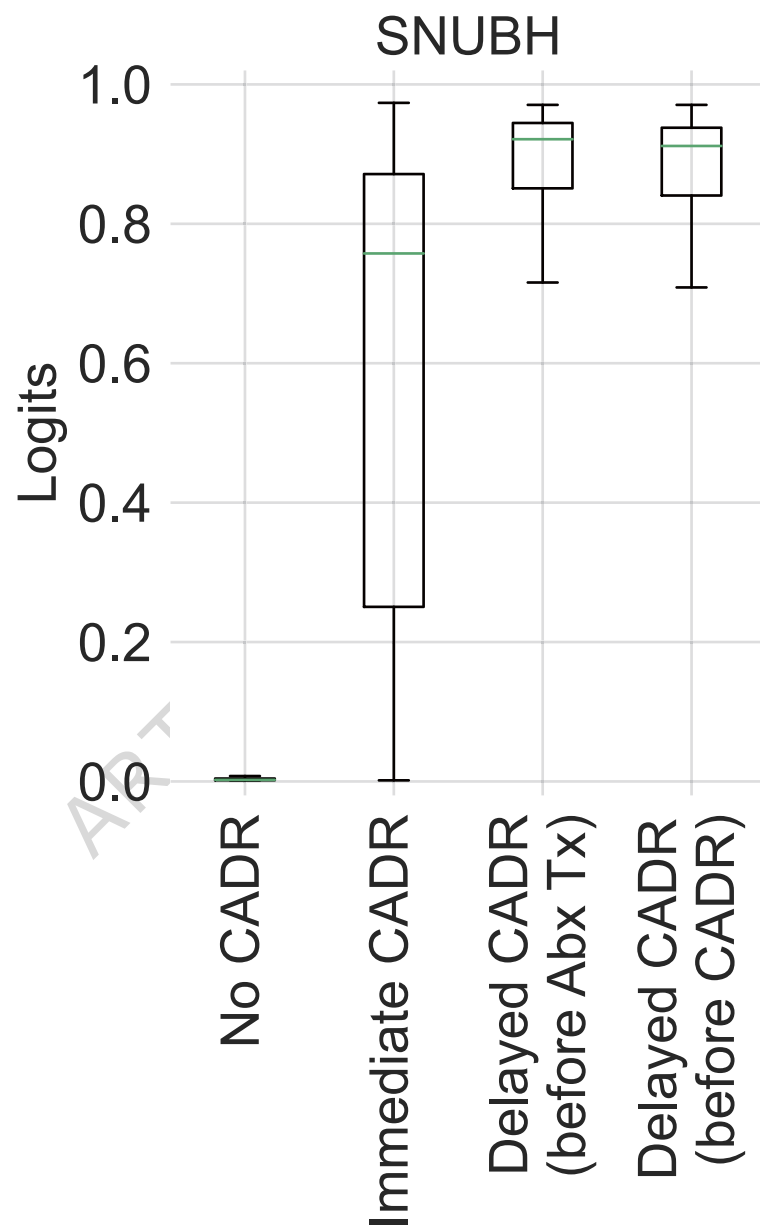
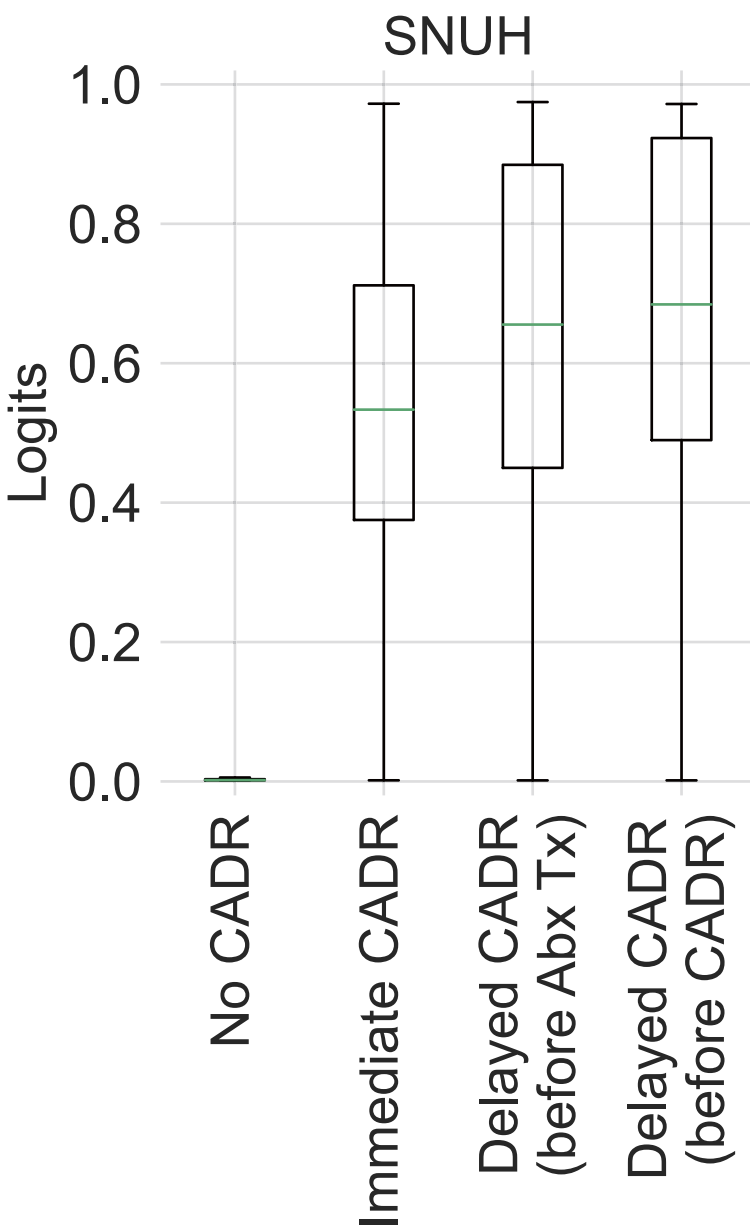
SNUBH

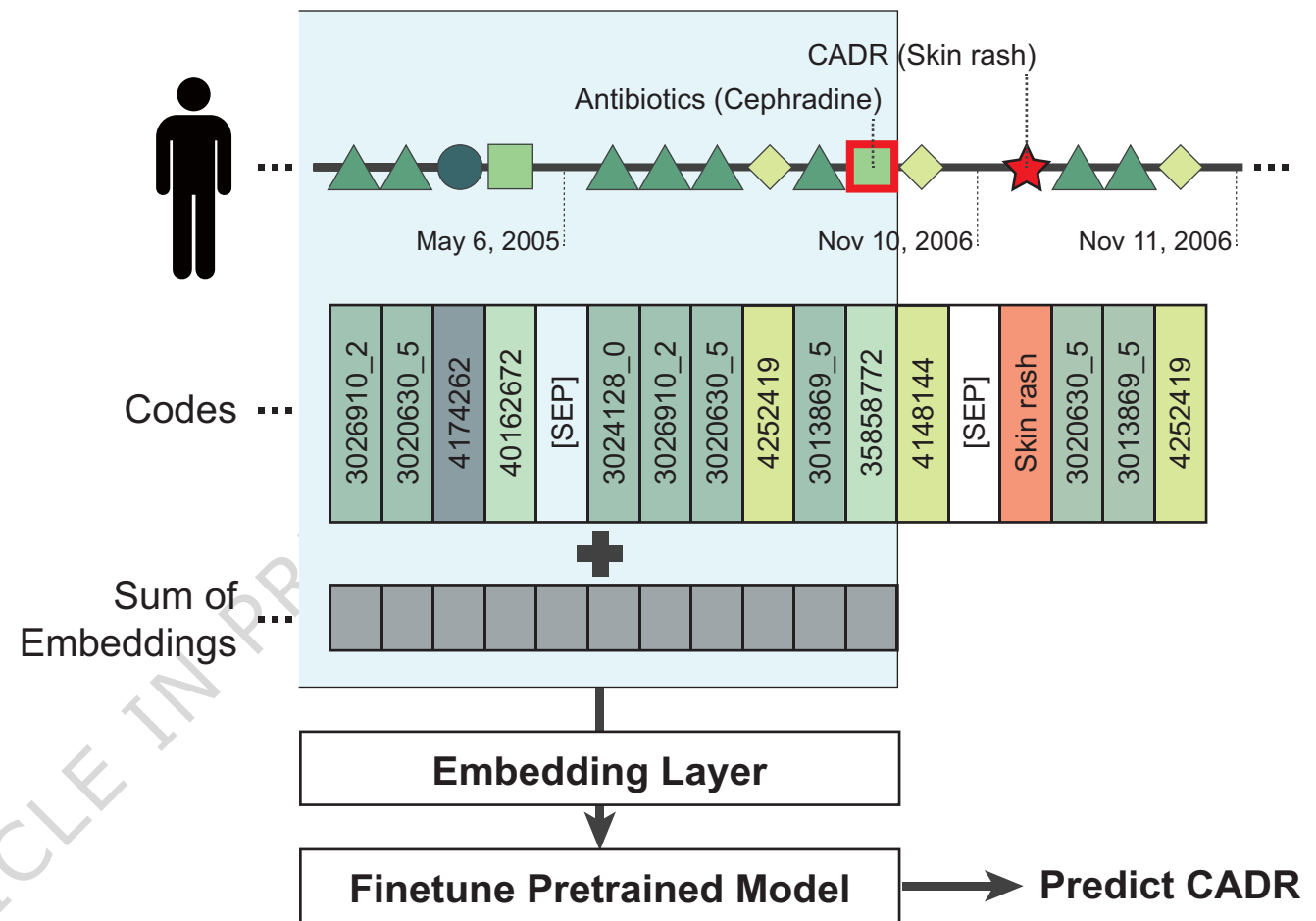
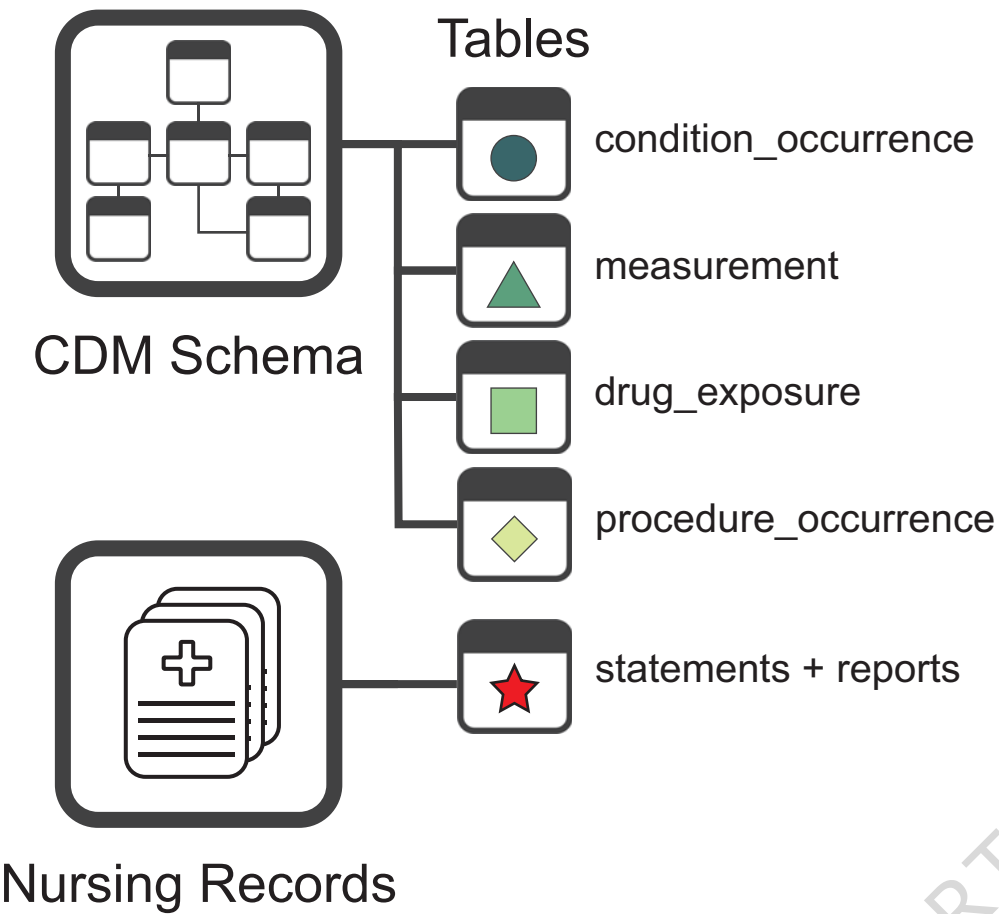
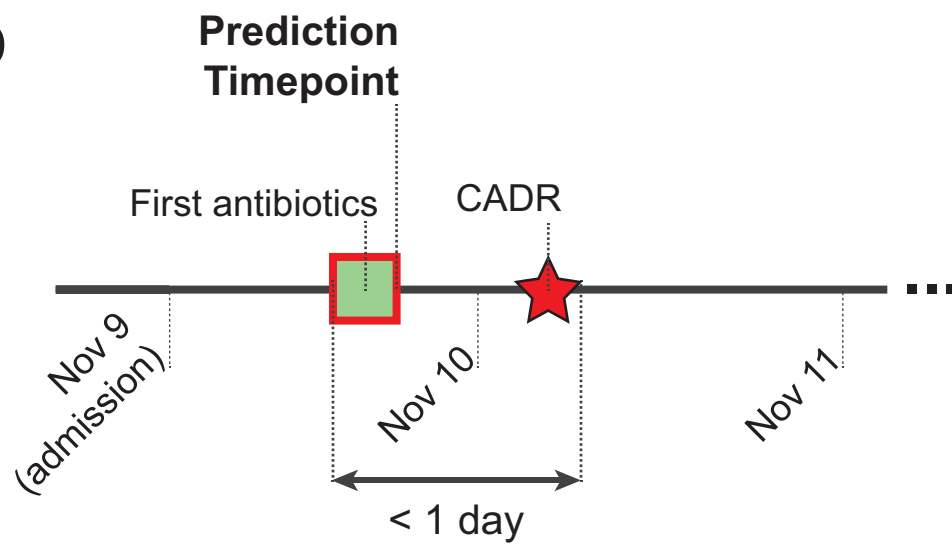


CNUH



Condition Drug Measurement Procedure



a**b****c**