scientific reports

Check for updates

OPEN

Natural language processing of electronic medical records identifies cardioprotective agents for anthracycline induced cardiotoxicity

Yoshimasa Kawazoe[1]✉, Masami Tsuchiya[2], Kiminori Shimamoto[1], Tomohisa Seki[3], Emiko Shinohara[1], Shuntaro Yada[4], Shoko Wakamiya[4], Shungo Imai[2], Eiji Aramaki[4] & Satoko Hori[2]

In this retrospective observational study, we aimed to investigate the potential of natural language processing (NLP) for drug repositioning by analyzing the preventive effects of cardioprotective drugs against anthracycline-induced cardiotoxicity (AIC) using electronic medical records. We evaluated the effects of angiotensin II receptor blockers/angiotensin-converting enzyme inhibitors (ARB/ACEIs), beta-blockers (BBs), statins, and calcium channel blockers (CCBs) on AIC using signals extracted from clinical texts via NLP. The study included 2935 patients prescribed anthracyclines at a single hospital, with concomitant prescriptions of ARB/ACEIs, BBs, statins, and CCBs. Upon propensity score matching, groups with and without these medications were compared, and expressions suggestive of cardiotoxicity, extracted via NLP, were considered as the outcome. The hazard ratios for ARB/ACEIs, BBs, statins, and CCBs were 0.58 [95% CI: 0.38–0.88], 0.71 [95% CI: 0.35–1.44], 0.60 [95% CI 0.38–0.95], and 0.63 [95% CI: 0.45–0.88], respectively. ARB/ACEIs, statins, and CCBs significantly suppressed AIC, whereas BBs did not demonstrate statistical significance, possibly due to limited statistical power. NLP-extracted signals from clinical texts reflected the known effects of these medications, demonstrating the feasibility of NLP-based drug repositioning. Further investigation is needed to determine if similar results can be replicated using electronic medical records from other institutions.

Keywords Natural Language processing, Electronic medical records, Drug repurposing, Anthracycline-induced cardiotoxicity

Drug repurposing, which involves discovering new uses for approved drugs, is an attractive approach for avoiding the "valley of death" in drug development[1–3]. As the side effects and pharmacokinetics of approved drugs are already known, repurposing is less likely to encounter development stagnation compared with traditional drug discovery processes, leading to significant time and cost savings. Retrospective analyses using large electronic medical records (EMRs) offer several advantages over randomized controlled trials (RCTs), including lower human and financial costs, reduced time to completion, and fewer ethical concerns. EMRs contain extensive data on patient phenotypes and outcomes, allowing for the identification of signals for drug repurposing[4]. Specifically, diagnostic codes, clinical test results, and prescription data, which are becoming more standardized and harmonized, have been actively used as outcomes for drug repurposing. Xu et al. identified an association between the diabetes drug metformin and improved cancer survival[5]. Kraus et al. demonstrated clinical benefits of adding palbociclib to endocrine therapy in male patients with metastatic breast cancer[6]. Shuey et al. showed that calcium channel blockers (CCBs) are promising candidates for lowering blood glucose levels and reducing cardiovascular disease risk[7]. Imai et al. reported that the sleep aid ramelteon reduces the risk of vancomycin-

[1]Artificial Intelligence and Digital Twin in Healthcare, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan. [2]Division of Drug Informatics, Keio University Faculty of Pharmacy, Tokyo, Japan. [3]Department of Healthcare Information Management, The University of Tokyo Hospital, Tokyo, Japan. [4]Division of Information Science, Graduate School of Science and Technology, Nara Institute of Science and Technology, Nara, Japan. ✉email: kawazoe@m.u-tokyo.ac.jp

induced nephrotoxicity[8]. Zamami et al. analyzed the FDA spontaneous adverse event (AE) reporting database and EMRs and found that simvastatin significantly mitigates oxaliplatin-induced peripheral neuropathy[9].

In pharmacological epidemiology, the use of EMRs is essential for achieving a comprehensive understanding of AEs and for conducting robust drug safety assessments and risk management. Notably, advancements in data mining technologies have enabled the extraction of meaningful insights from the abundance of data stored in EMRs. For example, applying machine learning and deep-learning models for AE detection helps in extracting valuable information not only from structured data, such as diagnostic codes and laboratory results, but also from unstructured free-text entries in clinical notes. Traditionally, pharmacoepidemiological studies using EMRs have primarily relied on AEs defined based on easily manageable structured data, such as diagnostic codes, laboratory test results for blood or urine samples, or their combinations. However, diagnostic codes are primarily intended for insurance claims rather than clinical diagnoses, resulting in low coverage of actual AEs[10,11]. Moreover, patient signs and symptoms suggestive of AEs, as well as healthcare provider observations, are generally recorded as free text. Consequently, the AE types that can be captured through structured data combinations are limited. Therefore, clinical texts containing rich information about AEs are crucial sources, and natural language processing (NLP) technologies for mining patient information are becoming increasingly important[12–14].

Accurate extraction of AEs from clinical texts presents several challenges: identifying AE-related symptoms and findings, determining whether they occurred in the patient, and normalizing the extracted varied expressions. In NLP, these challenges have been formulated and addressed as named entity recognition (NER), factuality analysis (FA), and entity normalization (EN) tasks, respectively.

The performance of these NLP tasks have improved dramatically with the introduction of bidirectional encoder representations from transformers (BERT)[15,16], which incorporate the transformer architecture introduced in 2017. BERT has significantly transformed NLP by employing a bidirectional approach that captures intricate contextual relationships within text. Unlike traditional models, BERT encodes each word by considering both its preceding and succeeding contexts, enabling the generation of contextualized representations that reflect the meaning of entire sentences. This capability addresses challenges, such as understanding ambiguity and resolving contextual nuances in language processing.

The ability of BERT to handle linguistic phenomena such as polysemy and homonymy is particularly valuable in medical contexts, where accurate interpretation of complex terminology is essential. By leveraging large-scale pretraining and its bidirectional architecture, BERT is highly effective in understanding specialized medical language and professional jargon. This has facilitated the development of domain-specific models, such as BioBERT[17] and other clinical BERT variants[18,19], which are pre-trained on biomedical and clinical texts, respectively.

Advancements in BERT-based models have significant implications for medical applications, particularly in extracting AEs from clinical texts[20–25]. By enhancing the ability to capture context and nuance in medical language, these models enable more reliable pharmacoepidemiological studies and contribute to better patient care. This includes improving drug repurposing efforts and enhancing AE monitoring systems, both of which are critical for advancing medical research and healthcare outcomes.

Given this background, we aimed to leverage longitudinal EMRs and combine signals extracted from clinical texts through NLP with other structured data to establish a framework for rapidly validating drug repurposing-related hypotheses. This study evaluates the effects of known cardioprotective agents against anthracycline (AC)-induced cardiotoxicity (AIC) as a validation of this framework.

With improving survival rates among patients with cancer, cardiotoxicity associated with cancer treatments has garnered increasing attention[26]. Specifically, ACs, widely used as standard treatments for breast cancer, ovarian cancer, and hematologic malignancies, exhibit dose-dependent cardiotoxicity[27]. Angiotensin II receptor blockers/angiotensin-converting enzyme inhibitors (ARB/ACEIs), beta-blockers (BBs), and HMG-CoA reductase inhibitors (statins) have been shown through multiple meta-analyses to exert cardioprotective effects against ACs[28–39]. Aldosterone antagonists (AAs) have also demonstrated cardioprotective effects against AIC in a previous RCT[40].

However, evidence regarding the efficacy of CCBs remains mixed. A pilot double-blind trial using prenylamine suggested potential mitigation of AIC by inhibiting calcium influx into myocardial cells[41], whereas a small RCT in acute myeloid leukemia showed no significant differences[42].

Therefore, we aimed to examine the effects of ARB/ACEIs, BBs, statins, AAs, and CCBs on AIC by exploring the utility and limitations of the proposed framework and the signals extracted from clinical texts through NLP. Table 1 summarizes the drug effects on AIC.

Results

Among the 44,502 patients hospitalized between 2004 and 2021, 2,935 who received at least one prescription of AC met the eligibility criteria and 240 met the exclusion criteria. Consequently, 2695 patients were included in the analysis. Among them, patients using ARB/ACEIs, BBs, statins, AAs, and CCBs, as well as those not using any of these medications, accounted for 168 (6.2%), 66 (2.4%), 135 (5.0%), 35 (1.3%), 256 (9.5%), and 2,182 cases (81.0%), respectively. The AA group was not analyzed due to an insufficient number of patients.

Since some patients were prescribed multiple medications, the total number of cases receiving each medication amounted to 2,842, exceeding the total of 2,695 patients included in the analysis. The details of single and combination therapy are presented below. Of the 168 patients using ARB/ACEIs, 105 were using ARB/ACEIs alone and 63 were using a combination with one or more of BBs, AAs, statins, or CCBs. Of the 66 cases using BBs, 37 were using BBs alone and 29 were using a combination with one or more of ARB/ACEIs, AAs, statins, or CCBs. Among the 135 patients using statins, 95 were using statins alone and 40 were using a combination with one or more of ARB/ACEIs, BBs, AAs, or CCBs. Of the 256 patients using CCBs, 181 were

	Effects on cardiotoxicity	Evidence (evidence level)	Sources
ARB/ACEI	Suppress	MetA (1a)	28–32
BB	Suppress	MetA (1a)	28, 29, 31–34
Statin	Suppress	MetA (1a)/RCT (1b)	31, 32, 35/36,37
	No significant effects	RCT (1b)	38, 39
AA	Suppress	RCT (1b)	40
CCB	Suppress	small RCT (1b)	41
	No significant effects	small RCT (1b)	42

Table 1. Summary of drug effects on anthracycline-induced cardiotoxicity. *AA* aldosterone antagonist, *ACEI* angiotensin-converting enzyme inhibitors, *ARB* angiotensin II receptor blockers, *BB* beta-blockers, statin, *CCB* calcium channel blockers, HMG-CoA reductase inhibitors, *MetA* meta-analysis, *RCT* randomized controlled trial.

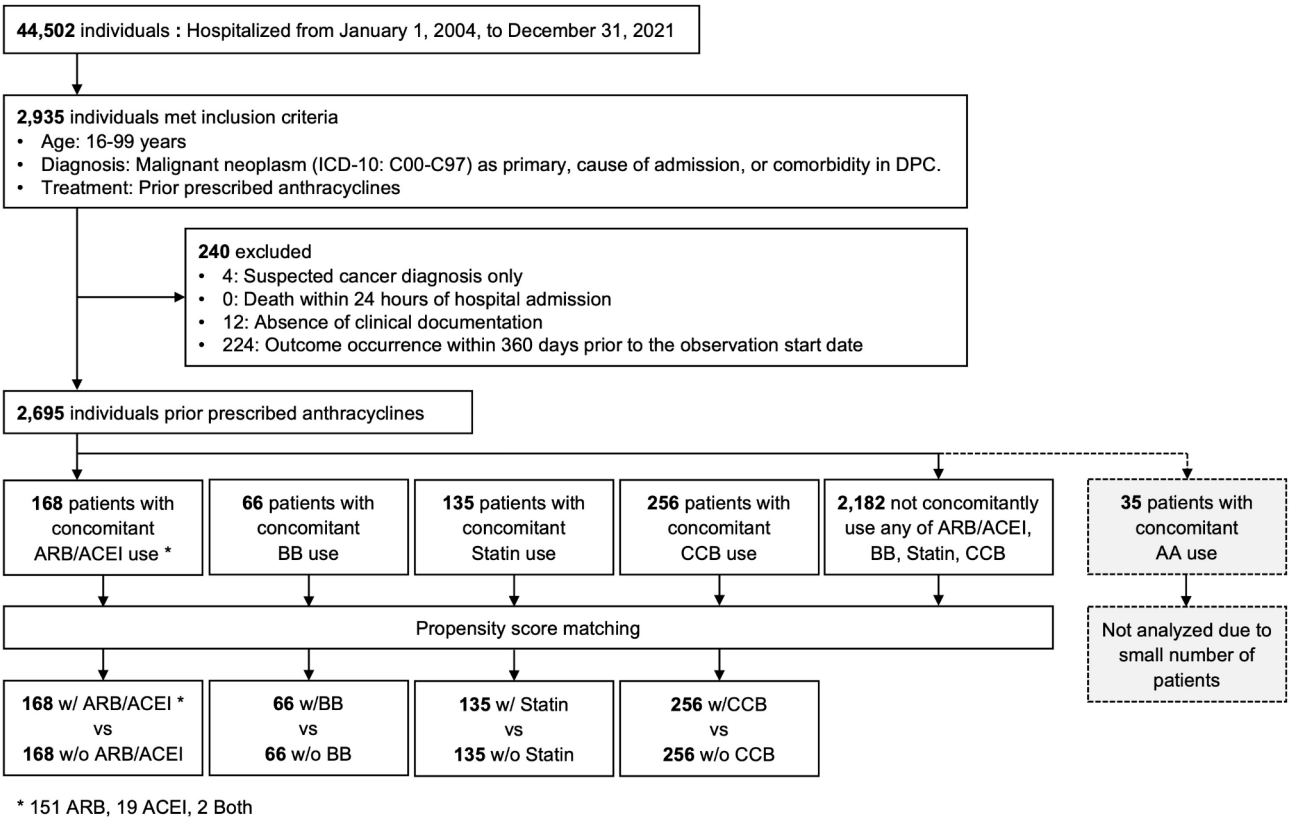

Fig. 1. Overview of the research cohort. Patients who were prescribed ARB/ACEI, BB, statin, or CCB concomitantly and those who were not were identified (AA was not analyzed owing to an insufficient number of patients). Based on propensity score matching, patients who received concomitant prescriptions were identified as the exposure group, and an equal number of patients who did not receive concomitant prescriptions were identified as the control group. *AA* aldosterone antagonist, *ACEI* angiotensin-converting enzyme inhibitors, *ARB* angiotensin II receptor blockers, *BB* beta-blockers, *CCB* calcium channel blockers, *DPC* diagnosis procedure combination, *ICD-10* International Classification of Diseases, 10th Revision; Statin, HMG-CoA Reductase Inhibitors, *w/* with, *w/o* without.

using CCBs alone and 75 were using a combination with one or more of ARB/ACEIs, BBs, AAs, or statins. Among the participants using ARB/ACEIs, 151 were on ARBs, 19 were on ACEIs, and two were using both medications. Figure 1 provides an overview of the research cohort.

The baseline characteristics of each cohort before propensity score matching (PSM) showed various differences (Table 2). The time to outcome ranged from 99.1 to 152.0 days, with the shortest durations observed in the AC with ARB/ACEI (w/ ARB/ACEI) and AC with CCB (w/ CCB) groups and the longest in the AC with BB (w/ BB) group.

Characteristics	Before propensity score matching				
	w/ARB/ACEI (*N* = 168)	w/BB (*N* = 66)	w/statin (*N* = 135)	w/CCB (*N* = 256)	Not use any of ARB/ACEI, BB, statin, CCB (*N* = 2182)
Time-to-event, mean, days (SD)	99.1 (96.4)	152.0 (104.0)	101.1 (98.1)	99.1 (88.5)	109.8 (92.3)
Demographics, n (%)					
Age at admission (< 65)	52 (31.0)	18 (27.3)	41 (30.4)	67 (26.2)	1344 (61.6)
Age at admission (65–79)	96 (57.1)	34 (51.5)	73 (54.1)	147 (57.4)	703 (32.2)
Age at admission (≥ 80)	20 (11.9)	14 (21.2)	21 (15.6)	42 (16.4)	135 (6.2)
Sex (Female)	69 (41.1)	25 (37.9)	69 (51.1)	96 (37.5)	1254 (57.4)
Cancer site, n (%)					
Digestive organs	4 (2.4)*	1 (1.5)*	1 (0.7)*	8 (3.1)*	42 (1.9)
Respiratory and intrathoracic organs	4 (2.4)*	4 (6.1)	8 (5.9)	10 (3.9)*	63 (2.9)
Bones and articular cartilage	2 (1.2)*	0 (0.0)*	0 (0.0)*	2 (0.8)*	26 (1.2)
Mesothelial and soft tissue	5 (3.0)*	2 (3.0)*	5 (3.7)*	9 (3.5)*	91 (4.2)
Breast	23 (13.7)	5 (7.6)	18 (13.3)	22 (8.6)	522 (23.9)
Female genital organs	10 (6.0)	4 (6.1)	9 (6.7)	14 (5.5)	182 (8.4)
Urinary tract	30 (17.9)	14 (21.2)	28 (20.7)	57 (22.3)	300 (13.7)
Metastatic	11 (6.5)	4 (6.1)	6 (4.4)*	10 (3.9)*	73 (3.3)
Lymphoid, hematopoietic and related tissue	89 (53.0)	36 (54.5)	64 (47.4)	135 (52.7)	959 (44.0)
Comorbidity, n (%)					
Diabetes	50 (29.8)	19 (28.8)	60 (44.4)	77 (30.1)	294 (13.5)
Ischemic heart disease	1 (0.6)*	3 (4.5)*	4 (3.0)*	1 (0.4)*	23 (1.1)
Heart failure	8 (4.8)*	5 (7.6)	4 (3.0)*	8 (3.1)*	60 (2.7)
Medication, n (%)					
Nitrogen mustard analogs	107 (63.7)	38 (57.6)	77 (57.0)	136 (53.1)	1,195 (54.8)
Pyrimidine analogs	24 (14.3)	5 (7.6)	7 (5.2)	26 (10.2)	449 (20.6)
Vinca alkaloids and analogs	80 (47.6)	36 (54.5)	65 (48.1)	131 (51.2)	756 (34.6)
Etoposide	5 (3.0)*	2 (3.0)*	4 (3.0)*	9 (3.5)*	60 (2.7)
Taxanes	4 (2.4)*	2 (3.0)*	5 (3.7)*	8 (3.1)*	50 (2.3)
Platinum compounds	20 (11.9)	5 (7.6)	19 (14.1)	33 (12.9)	228 (10.4)
Opioid	39 (23.2)	15 (22.7)	25 (18.5)	65 (25.4)	417 (19.1)
Diuretics (excluding AA)	35 (20.8)	13 (19.7)	21 (15.6)	49 (19.1)	276 (12.7)
ARB/ACEI	168 (100)	7 (10.6)	21 (15.6)	50 (19.5)	–
BB	7 (4.2)	66 (100)	11 (8.1)	19 (7.4)	–
AA	4 (2.4)	3 (4.5)	2 (1.5)	8 (3.1)	–
statin	21 (12.5)	11 (16.7)	135 (100)	23 (9.0)	–
CCB	50 (29.8)	19 (28.8)	23 (17.0)	256 (100)	

Table 2. Baseline characteristics of the study population before propensity score matching. In comparing the concomitant use groups of ARB/ACEI, BB, statin, and CCB, characteristics with a frequency < 5% in the cases (excluding ARB/ACEI, BB, AA, statin, CCB) were not used as covariates in the propensity score matching (indicated by *). *AA* aldosterone antagonist, *ACEI* angiotensin-converting enzyme inhibitors, *ARB* angiotensin II receptor blockers, *BB* beta-blockers, *CCB* calcium channel blockers.

The proportion of cases aged ≤ 65 years was highest in the AC without any (w/o Any) group at 61.6% (1,344 patients) and lowest in the w/ CCB group at 26.2% (67 cases). The proportion of cases aged 65–79 years was highest in the w/ CCB group at 57.4% (147 patients) and lowest in the w/o Any group at 32.2% (703 patients). The proportion of cases aged ≥ 80 years was highest in the w/ BB group at 21.2% (14 patients) and lowest in the w/o Any group at 6.2% (135 patients).

The groups with the highest proportion of female patients were the w/o Any group at 57.4% (1,254 patients) and the AC with statin (w/ statin) group at 51.1% (69 patients), followed by the w/ ARB/ACEI, w/ BB, and w/ CCB groups at 41.1% (69 patients), 37.9% (25 patients), and 37.5% (96 patients), respectively.

Regarding cancer sites, lymphoid, hematopoietic, and related tissues were the most common in all groups, with the highest proportion observed in the w/ BB group at 54.5% (36 patients), followed by the w/ ARB/ACEI, w/ CCB, w/ statin, and w/o Any groups at 53.0% (89 patients), 52.7% (135 patients), 47.4% (64 patients), and 44.0% (959 patients), respectively.

The urinary tract was the second most common site, with the highest proportion in the w/ CCB group at 22.3% (57 patients), followed by the w/ BB, w/ statin, w/ ARB/ACEI, and w/o Any groups at 21.2% (14 patients), 20.7% (28 patients), 17.9% (30 patients), and 13.7% (300 patients), respectively.

The breasts were the third most common site, with the highest proportion in the w/o Any group at 23.9% (522 patients), followed by the w/ ARB/ACEI, w/ statin, w/ CCB, and w/ BB groups at 13.7% (23 patients), 13.3% (18 patients), 8.6% (22 patients), and 7.6% (5 patients), respectively. Other cancer sites, including metastatic cases, accounted for < 10% in all groups.

Regarding comorbidities, the proportion of patients with diabetes was highest in the w/ statin group at 44.4% (60 patients), followed by the w/ CCB, w/ ARB/ACEI, and w/ BB groups at 30.1% (77 patients), 29.8% (50 patients), and 28.8% (19 patients), respectively, with the lowest proportion observed in the w/o Any group at 13.5% (294 patients). Ischemic heart disease was present in < 5% of all groups. Heart failure was observed in 7.6% (5 patients) of the w/ BB group but was < 5% in all other groups.

Regarding concomitant anticancer drug use, nitrogen mustard analogs were the most commonly used in all groups, with the highest usage observed in the w/ ARB/ACEI group at 63.7% (107 patients) and 53.1–57.6% in the other groups. Vinca alkaloids and analogs were the second most commonly used, with the highest usage in the w/ BB group at 54.5% (36 patients) and 34.6–51.2% in the other groups. Platinum compound usage ranged from 7.6 to 14.1%, whereas pyrimidine analog usage ranged from 5.2 to 20.6%. Etoposide and taxane usage remained below 5% in all groups. Opioid usage ranged from 18.5 to 25.4%, whereas diuretic use (excluding AAs) ranged from 12.7 to 20.8%.

Regarding other cardioprotective drug use, ARB/ACEI usage was highest in the w/ CCB group at 19.5% (50 patients), followed by the w/ statin (15.6%, 21 patients) and the w/ BB (10.6%, 7 patients) groups. BB usage was highest in the w/ statin group at 8.1% (11 patients), followed by the w/ CCB (7.4%, 19 patients) and the w/ ARB/ACEI (4.2%, 7 patients) groups. AA usage was highest in the w/ BB group at 4.5% (3 patients), followed by the w/ CCB (3.1%, 8 patients), the w/ ARB/ACEI (2.4%, 4 patients) and the w/ statin (1.5%, 2 patients) groups. Statin usage was highest in the w/ ARB/ACEI group at 12.5% (21 patients), followed by the w/ BB (16.7%, 11 patients) and the w/ CCB (9.0%, 23 patients) groups. CCB usage was highest in the w/ ARB/ACEI group at 29.8% (50 patients), followed by the w/ BB group at 28.8% (19 patients) and the w/ statin group at 17.0% (23 patients).

For each cohort, PSM identified an equal number of unexposed patients for the exposed group (Supplementary Tables S1–4). The mean (SD) days to outcome after PSM were 99.1 (96.4) versus 97.8 (91.4) days for the w/ ARB/ACEI group, 152.0 (104.0) versus 114.3 (92.0) days for the w/ BB group, 101.1 (98.1) versus 97.8 (75.0) days for the w/ statin group, and 99.1 (88.5) versus 116.0 (97.2) days for the w/ CCB group. No significant differences in days to outcome were observed between any groups. Additionally, no significant differences were observed in the mean number of hospital visits, including both inpatient and outpatient visits, between the groups.

In comparisons where covariates had an absolute standardized difference (ASD) exceeding 10%, multivariate Cox proportional hazards models were employed to estimate adjusted hazard ratios (aHRs), controlling for these imbalanced covariates (Supplementary Tables S5–8). The covariates exceeding 10% ASD were breast cancer (11.4%) and metastatic cancer (20.2%) in the ARB/ACEI comparison; age < 65 years (17.7%), age ≥ 80 years (14.0%), female sex (18.2%), respiratory and intrathoracic cancer (14.4%), breast cancer (15.1%), and diuretic (excluding AAs) use (10.8%) in the BB comparison; and ARB/ACEI use (13.0%) in the statin comparison; and opioid use (13.1%) in the CCB comparison.

Table 3 summarizes the hazard ratios (HRs) for AIC. The crude HRs (cHRs) for the w/ ARB/ACEI, w/ BB, w/ statin, and w/ CCB groups were 0.55 [95% confidence interval (CI): 0.36–0.84] ($p < 0.01$), 0.78 [95% CI: 0.39–1.53] ($p = 0.47$), 0.60 [95% CI: 0.38–0.94] ($p < 0.05$), and 0.63 [95% CI: 0.45–0.88] ($p < 0.01$), respectively. The adjusted HRs (aHRs) were 0.58 [95% CI: 0.38–0.88] ($p < 0.05$) for the w/ ARB/ACEI group, 0.71 [95% CI: 0.35–1.44] ($p = 0.34$) for the w/ BB group, 0.60 [95% CI: 0.38–0.95] ($p < 0.05$) for the w/ statin group, and 0.63 [95% CI: 0.45–0.88] ($p < 0.01$) for the w/ CCB group. Both cHR and aHR demonstrated that ARB/ACEI, statins,

	Before propensity score matching		After propensity score matching			
	w/ concomitant (N with event)	w/o concomitant (N with event)	w/ concomitant (N with event)	w/o concomitant (N with event)	Crude HR [95%CI] (p-value)	Adjusted HR [95%CI] (p-value)
ARB/ACEI	*N* = 168	*N* = 2527	*N* = 168	*N* = 168	0.55 [036–0.84] (< 0.01)	0.58 [0.38–0.88] (< 0.05)
	(35, 20.8%)	(360, 14.2%)	(35, 20.8%)	(58, 34.5%)		
BB	*N* = 66	*N* = 2,629	*N* = 66	*N* = 66	0.78 [0.39–1.53] (= 0.47)	0.71 [0.35–1.44] (= 0.34)
	(15, 22.7%)	(380, 14.5%)	(15, 22.7%)	(19, 28.8%)		
Statin	*N* = 135	*N* = 2560	*N* = 135	*N* = 135	0.60 [0.38–0.94] (< 0.05)	0.60 [0.38–0.95] (< 0.05)
	(30, 22.2%)	(365, 14.3%)	(30, 22.2%)	(48, 35.6%)		
CCB	*N* = 256	*N* = 2439	*N* = 256	*N* = 256	0.63 [0.45–0.88] (< 0.01)	0.63 [0.45–0.88] (< 0.01)
	(56, 21.9%)	(339, 13.9%)	(56, 21.9%)	(88, 34.4%)		

Table 3. Summary of hazard ratios for AIC after propensity score matching. *ACEI* angiotensin-converting enzyme inhibitors, *AIC* anthracyclines-induced cardiotoxicity, *ARB* angiotensin II receptor blockers, *BB* beta-blockers, *CCB* calcium channel blockers, *CI* confidence interval, *HR* hazard ratio, statin, HMG-CoA reductase inhibitors; *w/* with, *w/o* without, *SD* standard deviation, statin, HMG-CoA reductase inhibitors, *w/* with, *w/o* without.

and CCB significantly reduced AIC risk. Although BB showed a tendency toward AIC suppression, it did not reach statistical significance.

Figure 2 presents the cumulative incidence curves and the log-log transformed cumulative incidence curves. In the log-rank test, all comparisons except for BB showed a significant reduction in cHRs in the w/ groups. In the ARB/ACEI comparison, no intersection was observed in the log-log transformed cumulative incidence curves, confirming the proportional hazards assumption. However, an intersection was observed around day 20 (e^3) for the statin comparison and around day 30 ($e^{3.4}$) for the CCB comparison.

The power analysis of the adjusted model for the groups with and without ARB/ACEI ($N=336$), with and without BB ($N=132$), with and without statin ($N=270$), and with and without CCB ($N=512$) yielded power values of 75.7%, 17.0%, 61.7%, and 80.2%, respectively.

Regarding the first sensitivity analysis, HRs according to the cumulative number of AC prescriptions showed results consistent with the primary analysis across all comparisons (Supplementary Tables S9–13). In contrast, the log-log transformed cumulative incidence curves based on the number of AC prescriptions did not intersect in any comparisons, including those with statin, confirming the proportional hazards assumption except for BB comparison (Fig. 3). As for the second sensitivity analysis, even when the concomitant medication period was extended to 90 days or shortened to 30 days, the HRs remained consistent with the primary analysis across all comparisons (Supplementary Tables S14–24, Supplementary Fig. S1, S2). Regarding the third sensitivity analysis, proton pump inhibitors (PPIs) did not show a significant reduction in HRs, confirming their role as a negative control (Supplementary Tables S25–27, Supplementary Fig. S3).

Discussion

This retrospective longitudinal observational study evaluated the effects of cardioprotective drugs on AIC using AEs extracted from clinical text via NLP as signals. Although AEs have previously been extracted from medical text using NLP[20–25], they have not been evaluated as time-to-event outcomes for drug repositioning. Our

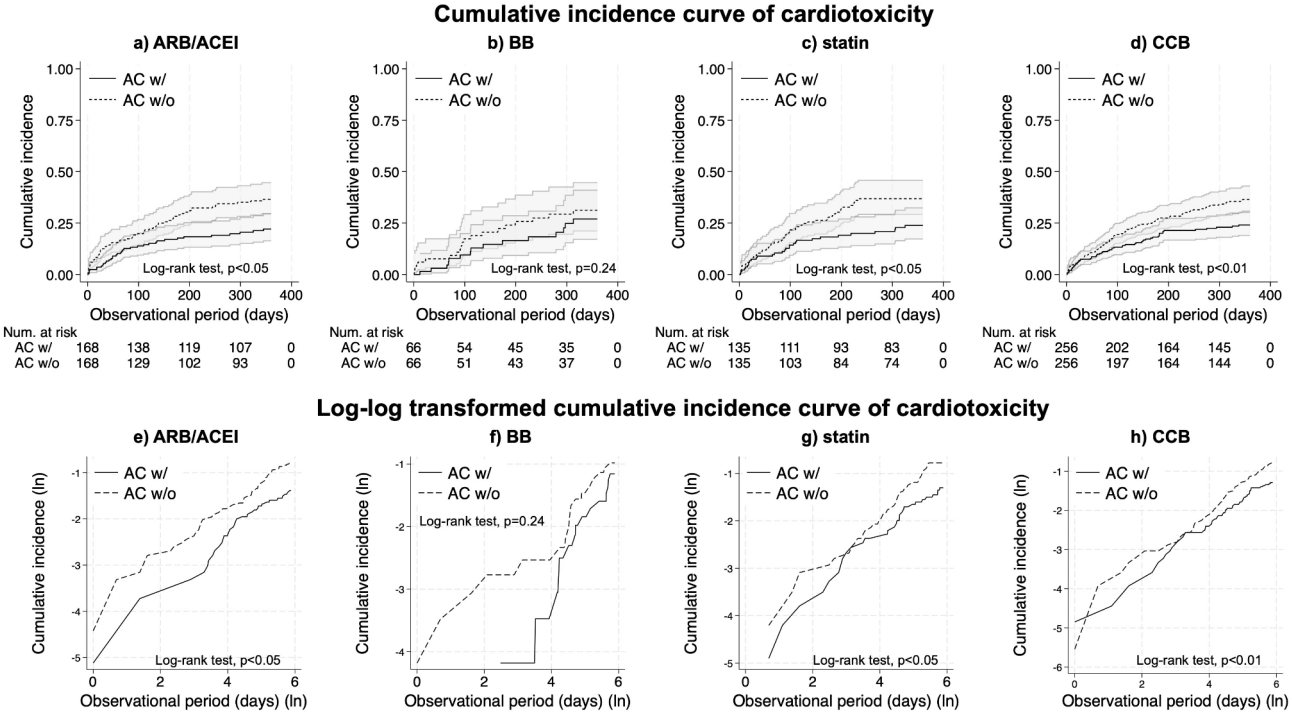

Fig. 2. Cumulative incidence curves and log-log transformed cumulative incidence curves for 12-month freedom outcomes.Top: Cumulative incidence curve. (**a**) Comparison between the ARB/ACEI concomitant and non-concomitant groups. (**b**) Comparison between the BB concomitant and non-concomitant groups. (**c**) Comparison between the statin concomitant and non-concomitant groups. (**d**) Comparison between the CCB concomitant and non-concomitant groups.The horizontal axis represents the number of days, while the vertical axis denotes the cumulative cardiotoxicity incidence. Bottom: Log-log transformed cumulative incidence curve. (**e**) Comparison between the ARB/ACEI concomitant and non-concomitant groups. (**f**) Comparison between the BB concomitant and non-concomitant groups. (**g**) Comparison between the statin concomitant and non-concomitant groups. (**h**) Comparison between the CCB concomitant and non-concomitant groups. The horizontal axis represents the natural logarithm of the number of days, while the vertical axis denotes the natural logarithm of the cumulative cardiotoxicity incidence. For the horizontal axis, e^2 corresponds to 7.4 days, e^3 to 20.1 days, e^4 to 54.6 days, e^5 to 148.4 days, and e^6 to 403.0 days. *AC* anthracyclines, *ARB* angiotensin II receptor blockers, *ACEI* angiotensin-converting enzyme inhibitors, *BB* beta-blockers, Statins, HMG-CoA Reductase Inhibitors, *CCB* calcium channel blockers, *w/* with, *w/o* without.

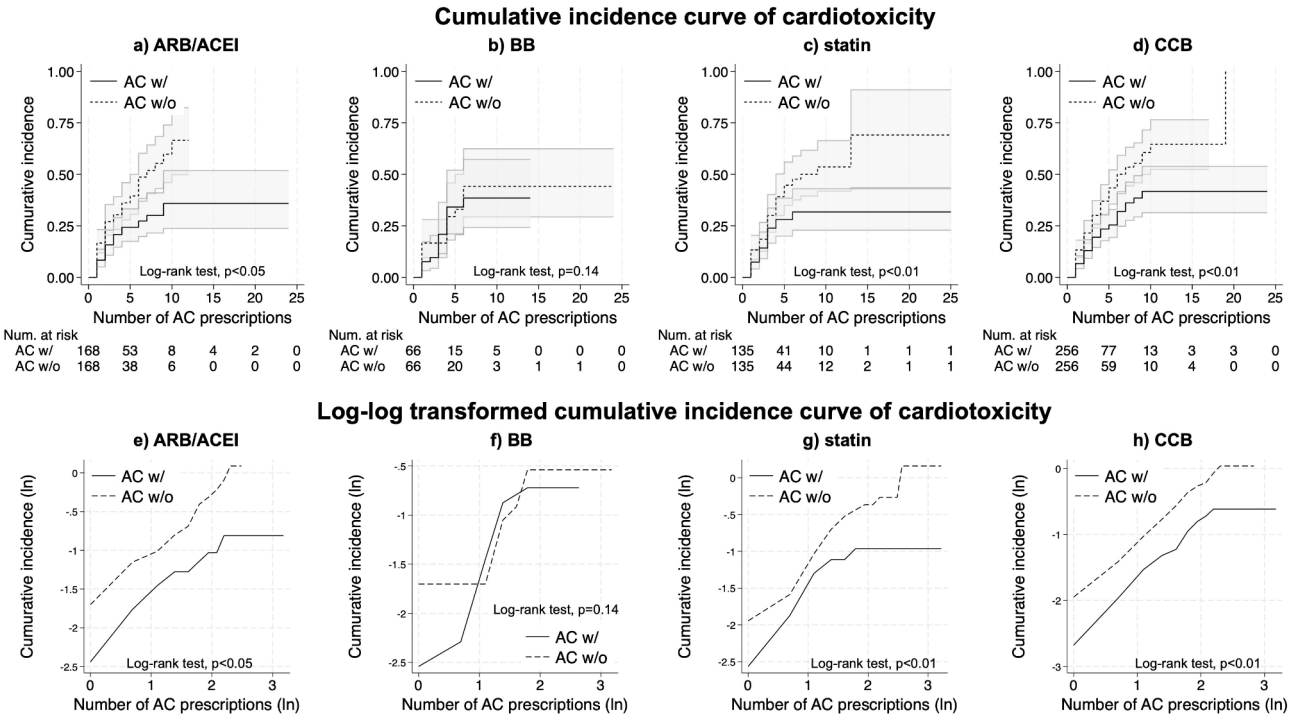

Fig. 3. Cumulative incidence curves and log-log transformed cumulative incidence curves for 12-month freedom outcomes based on the number of AC prescriptions. Top: Cumulative incidence curve. (**a**) Comparison between the ARB/ACEI concomitant and non-concomitant groups. (**b**) Comparison between the BB concomitant and non-concomitant groups. (**c**) Comparison between the statin concomitant and non-concomitant groups. (**d**) Comparison between the CCB concomitant and non-concomitant groups. The horizontal axis represents the number of days, while the vertical axis denotes the cumulative cardiotoxicity incidence. Bottom: Log-log transformed cumulative incidence curve. (**e**) Comparison between the ARB/ACEI concomitant and non-concomitant groups. (**f**) Comparison between the BB concomitant and non-concomitant groups. (**g**) Comparison between the statin concomitant and non-concomitant groups. (**h**) Comparison between the CCB concomitant and non-concomitant groups. The horizontal axis represents the natural logarithm of the number of prescriptions, while the vertical axis denotes the natural logarithm of the cumulative cardiotoxicity incidence. For the horizontal axis, e^1 corresponds to approximately 2.7 prescriptions, e^2 to 7.4 prescriptions, and e^3 to 20.1 prescriptions. *AC* anthracyclines, *ARB* angiotensin II receptor blockers, *ACEI* angiotensin-converting enzyme inhibitors, *BB* beta-blockers, Statin, HMG-CoA Reductase Inhibitors, *CCB* calcium channel blockers, *w/* with, *w/o* without.

results confirmed that ARB/ACEIs and statins, which reportedly have AIC-suppressive effects in meta-analyses, significantly suppressed AIC. Although only small-scale RCTs with conflicting results have been reported for CCBs, this study demonstrated that CCBs also significantly suppressed AIC. However, while BBs showed a tendency toward AIC suppression, they did not demonstrate a significant difference.

Post-hoc evaluation of statistical power revealed that the adjusted models for CCBs had power exceeding 80%, suggesting that the sample sizes for these models were adequate. However, the adjusted model for BBs demonstrated a low power at 17.0%, implying an 83.0% probability of a Type II error (false negative), quantitatively confirming that the lack of statistical significance was likely due to insufficient sample size. Additionally, the power for the ARB/ACEI and statin models was low at 75.7% and 61.7%, respectively, indicating a possibility of inadequate sample size. Although these two models demonstrated significant suppression of AIC, further validation with larger sample sizes is necessary to estimate more accurate HRs. AAs could not be analyzed due to an insufficient number of cases.

In the first sensitivity analysis, we evaluated the dose dependency of AIC by estimating HRs based on the frequency of AC prescriptions as a proxy for dosage due to database limitations. The results were consistent with the primary analysis for all cardioprotective drugs. The log-log transformed cumulative incidence curves (Fig. 3, lower panel) showed parallel lines except for the BB comparison, supporting the proportional hazards assumption and linking AIC occurrence to prescription frequency. However, the relationship between cardioprotective drug dosage or prescription frequency and HR was not assessed. While the studied cardioprotective drugs are often prescribed long-term, we cannot exclude the possibility of discontinuation or new prescriptions in both exposed and unexposed groups after the observation start date, which limits our study.

The second sensitivity analysis explored varying pre-exposure periods for cardioprotective drugs. Shortening this period risks misclassifying exposed cases as unexposed, while lengthening it risks the reverse. Since most prescriptions were estimated to last no more than 90 days, a 60-day period was chosen for the primary analysis,

with 30-day and 90-day results included for sensitivity analyses. The consistency across these periods confirmed the robustness of our findings.

In the third sensitivity analysis, PPIs were used as a negative control to confirm that the observed associations were not due to confounding, thereby supporting the validity of our results. These findings indicate that even signals extracted from clinical text using NLP reflect known drug effects, suggesting the feasibility of drug repositioning utilizing NLP. Although this study used a single-center cohort and requires validation in multicenter cohorts, the results suggest that signals extracted through NLP may facilitate the discovery of novel drug efficacies.

One of the most widely accepted mechanisms of AIC is the increase in reactive oxygen species. ARB/ACEIs and BBs prevent AIC by activating myocardium-protective signaling pathways and improving cardiac remodeling, while statins mitigate AIC through their pleiotropic effects by reducing inflammation and oxidative stress[43]. Gao et al. demonstrated in a meta-analysis that ARB/ACEI and BB significantly improve left ventricular ejection fraction (LVEF)[28]. Lewinter et al. reported in a meta-analysis that BB and ARB/ACEI mitigate LVEF decline during trastuzumab and AC therapy in women with breast cancer[29]. Kalam et al. found in a meta-analysis that ARB/ACEI, BB, and statins reduce the relative risk of AIC by 89%, 69%, and 69%, respectively[31]. Mir et al. reported in a network meta-analysis that ACEI, BB, and statins significantly reduce AIC, whereas ARB/ACEI had no significant effect[32]. In a meta-analysis, Titus et al. reported that statin use in patients with lymphoma or breast cancer reduces cardiotoxicity incidence by 54%[35]. In a double-blind RCT, Neilan et al. reported a cardiotoxicity incidence of 9% in the atorvastatin group compared to 22% in the placebo group, indicating a 13% absolute risk reduction[37]. In this study, the use of ARB/ACEIs and statins were confirmed to significantly suppress AIC, consistent with these previous findings.

In Japan, ARBs and CCBs are widely used antihypertensive agents, whereas ACEI and BB usage is relatively low[44,45]. In this study, the number of patients receiving BBs was relatively limited (66 cases), and as aforementioned, the small sample size likely resulted in reduced statistical power, potentially leading to the lack of significant difference observed. Regarding statins and CCBs, both cHR and aHR showed significant AIC suppression; however, visual confirmation of the proportional hazards assumption revealed an intersection in the curves around day 20 (e^3) and day 30 ($e^{3.4}$), respectively (Fig. 2). Conversely, log-log transformed cumulative incidence curves based on the cumulative number of AC prescriptions maintained the proportional hazards assumption across all groups, including statins and CCBs (Fig. 3). This discrepancy may be attributed to the lack of available regimen information in this study, which precluded adjustment for differences in AC administration intervals through PSM, leading to varied administration intervals between the two groups in the statin and CCB comparisons. These variations could have contributed to the observed curve intersection.

In patients receiving CCBs, AIC was significantly suppressed. Two small-scale RCTs have reported on CCBs. Milei et al. evaluated the effect of prenylamine in 26 adults with solid tumors undergoing doxorubicin chemotherapy; one case of cardiomyopathy occurred in the control group ($n=13$), whereas none occurred in the intervention group ($n=13$), suggesting that concomitant prenylamine administration may reduce ADM cardiotoxicity[41]. Conversely, Kraft et al. studied low-dose verapamil in adult patients with acute myeloid leukemia; one case of heart failure occurred in the control group ($n=17$), but none in the intervention group ($n=13$), showing no significant difference[42]. The mechanism by which CCBs affect AIC remains unclear. Monti et al. hypothesized that controlling excessive calcium influx into cardiomyocytes would have a protective effect against AIC and evaluated the effect of doxorubicin and CCB combination therapy using rat cardiomyocytes. However, contrary to the hypothesis, the concomitant use of CCBs was found to enhance AIC[46]. Similarly, Santostasi et al. reported that AIC induced by doxorubicin and daunorubicin might be exacerbated by CCBs in rat cardiomyocytes[47]. Although the effect of CCBs on AIC remains uncertain, the potential for CCBs to worsen AIC in animal models suggests that conducting RCTs in humans may be ethically challenging. Therefore, observational studies using EMRs are becoming increasingly important for verifying clinically contentious drug effects.

Regarding NLP, the BERT model embedded in MedNERN-CR-JA (MedNERN) used in this study was pre-trained on Japanese Wikipedia and fine-tuned on medical texts to achieve high-performance NER in medical contexts. However, as with any machine learning model, a certain degree of false positives and false negatives could have affected the accuracy of cardiotoxicity detection and time-to-event outcomes. One limitation of this study is that we did not conduct an error analysis of the NLP system and, therefore, have not quantitatively demonstrated the rates of under-detection or over-detection of outcomes due to NLP. However, the average number of visits did not differ between the compared groups (Supplementary Tables S1–4), suggesting no significant difference in the opportunity for recording cardiotoxicity in the EMRs. Furthermore, since the same NLP tool was used to extract outcomes for both groups, false positives and false negatives likely occurred at similar rates in both groups. Therefore, the impact of NLP errors on the outcomes is considered limited. Supporting this view, Lan and Turchin suggested that the impact of NLP errors on downstream analyses in epidemiological studies using NLP-derived data is minimal[48]. However, the quantity and quality of recorded text per visit may vary, potentially leading to missed or over-detected outcomes owing to NLP errors, which remains a limitation of this study. While our approach has general applicability, it is based on single-center results. Further investigation is thus needed to determine whether similar results can be obtained from EMRs at other institutions with different medical and documentation practices.

Spontaneous reporting systems (SRS), such as the FDA Adverse Event Reporting System, are valuable resources for post-marketing surveillance and drug repositioning. Signal detection using SRS typically employs odds ratios calculated based on drug use and specific AE reporting. While SRS provides large-scale, comprehensive AE data for various signal detections, its interpretation is limited by potential underreporting biases and the lack of denominator information for incidence calculations[49]. Furthermore, SRS cannot account for covariates, potentially leading to detection errors due to patient background imbalances. Conversely,

distributed EMR-derived databases, such as the Sentinel Initiative[50] and MID-NET[51], offer relatively large-scale, detailed patient information. However, they require defining signals through combinations of diagnostic codes and laboratory results, potentially missing AEs not registered as International Classification of Diseases, 10th Revision (ICD-10) codes in routine clinical practice. The single-institution EMRs utilized in this study, although small-scale, contains patient background information and medical text, enabling risk estimation for AEs not registered as ICD-10 codes while adjusting for patient characteristics. By treating AEs as time-to-event outcomes, this approach can incorporate censored cases into HR calculations, allowing evaluation of long-term treatment effects. Recent advancements in NLP using large language models are expected to enhance the accurate extraction of rare signals recorded in clinical texts[52]. Applying this approach with improved NLP capabilities could enable more rapid screening of novel drug effects across a broader range of post-marketing pharmaceuticals.

In conclusion, using signals extracted from clinical text via NLP tools as outcomes, we conducted a retrospective observational study to investigate the effects of ARB/ACEIs, BBs, statins, and CCBs on AIC. ARB/ACEIs and statins demonstrated significant known inhibitory effects, whereas BBs did not reach statistical significance, possibly due to limited statistical power. Despite conflicting reports regarding CCBs, this study demonstrated that CCBs significantly suppressed AIC. These findings suggest that signals extracted from clinical text using NLP can reflect known drug effects, supporting the feasibility of drug repositioning utilizing NLP. Furthermore, signals extracted through NLP may facilitate the discovery of novel drug efficacies. However, this study has several limitations. The relationships between the dosage or prescription duration of cardioprotective drugs and HRs were not considered. Additionally, differences in NLP errors between the groups, depending on the quantity and qualitative content of recorded texts, may have impacted the outcomes. Finally, as this study utilized a single-center cohort, further investigation is needed to determine whether similar results can be obtained from EMRs at other institutions.

Methods

Study design

A retrospective cohort study was conducted using the EMRs of a single institution, the University of Tokyo Hospital. Data collection and all experiments were approved by the institutional review board at the University of Tokyo and University of Tokyo Hospital (approval number: 2022251NI). Informed consent was obtained through an opt-out approach, which was approved by the institutional review board. Detailed information about the study was made publicly available on the hospital's clinical research center and department websites, providing easy access for patients and their families. The opt-out process, including the right to refuse participation and the method for doing so, was clearly explained. The dataset analyzed during the current study is not publicly available because of contracts with the hospital providing data to the database. All experiments were conducted under the relevant ethical guidelines and regulations. This study is reported following the Strengthening the Reporting of Observational Studies in Epidemiology reporting guidelines.

Settings

We utilized the diagnosis procedure combination (DPC) data from patients admitted to the University of Tokyo Hospital aged > 18 years from January 2004 to December 2021. In Japan, a DPC-based payment system has been implemented in acute care hospitals nationwide since 2003[53]. The DPC data include patient demographics, primary diagnoses, diseases leading to hospitalization, comorbidities at admission, and complications during hospitalization, all coded according to the ICD-10. The sensitivity of diagnoses registered in the DPC is 50–80%, while the specificity exceeds 96%[54], making it widely used in clinical epidemiological research. Along with DPC data, we utilized discharge summaries, prescription orders, injection orders, physician's notes, and nursing notes. We obtained and used comprehensive information covering the entire patient course, including both inpatient and outpatient data. The pharmaceutical types were analyzed by mapping the domestic standard drug codes included in the prescription and injection orders to the anatomical therapeutic chemical (ATC) classification codes. DPC, prescription orders, and injection orders were structured data, whereas physician's notes, nursing notes, and discharge summaries were recorded as free text. Therefore, we employed NLP tools to extract instances of cardiotoxicity from these clinical texts, as described in the outcome and NLP tool section.

Patients

The study included patients aged 16–99 years who had received at least one prescription of AC, regardless of the dosage, and had a primary diagnosis, admission-related disease, or comorbid condition listed under ICD-10 codes for malignant neoplasms (C00-C97) in the DPC. The age range was set to exclude individuals aged < 16 years due to challenges in obtaining consent from legal representatives using the opt-out method, and those aged ≥ 100 years to minimize identification risks. This criterion was approved by the ethics review board. All cancer stages were included without restrictions on cancer progression or specific classifications. Furthermore, patients were included regardless of their treatment history, encompassing those who had undergone surgical interventions, radiotherapy, or any other cancer treatment modalities. The exclusion criteria were as follows: (1) only a suspected cancer diagnosis, (2) death within 24 h of admission, (3) absence of clinical texts, and (4) experiencing the outcome within 360 days prior to the observation period start.

Exposure/comparison

We defined the concomitant medication period as 60 days prior to the initial AC administration date. Among eligible patients, those who received prescriptions for ARB/ACEIs, BBs, statins, AAs, or CCBs during this period were classified as the exposure group, designated as w/ ARB/ACEIs, w/ BBs, w/ statins, w/ AAs, and w/ CCBs, respectively. Owing to an insufficient number of cases, AA was excluded from further analysis. Patients who did

not receive prescriptions for ARB/ACEIs, BBs, statins, or CCBs during the same period were identified as w/o Any. The non-exposure groups were identified using PSM[55] from patients who did not receive prescriptions for ARB/ACEIs, BBs, statins, or CCBs during the concomitant period. These groups were designated as AC without ARB/ACEIs (w/o ARB/ACEI), AC without BBs (w/o BB), AC without statins (w/o statin), and AC without CCBs (w/o CCB). Comparisons were made between w/ ARB/ACEI and w/o ARB/ACEI, w/ BB and w/o BB, w/ statin and w/o statin, and w/ CCB and w/o CCB. Patients who received combination drugs composed of different medication classes were treated as having been prescribed each component separately. Figure 4 illustrates the definition of the concomitant medication period and the timeline of the observation period in this study. Supplementary Table S28 provides the definitions for each medication class based on ATC codes.

Outcome and NLP tool

The observation period commenced on the date of initial AC administration, and the outcome was defined as the appearance of textual expressions indicating cardiotoxicity in the clinical notes within 360 days following this date. We extracted the cardiotoxicity expressions from physician's notes, nursing notes, and discharge summaries using MedNERN[56], an NLP tool developed by the co-authors. This tool performs NER using a Japanese BERT model pre-trained on 17 million Japanese Wikipedia sentences and fine-tuned on approximately 2,000 Japanese medical text corpus sentences. Then, it normalizes the extracted named entities to terms in an internal dictionary for EN. NER assigns 12 types of named entity classes, including disease names (symptoms and findings) and temporal expressions. Disease name classes are further categorized by four factuality attributes: Positive, Negative, Suspicious, and General. Positive indicates confirmed presence or observation of the named entity, while Negative denotes its absence. Suspicious is used for differential diagnoses, and General for general knowledge about diseases. The public version of MedNERN includes an ICD-10 dictionary for EN; however, we created and used a new normalization dictionary that includes cardiotoxicity and multiple AEs, and comprises surface and normalized forms. For example, "ecreased echocardiographic function" (surface form) corresponds to "Cardiac dysfunction" (normalized form), while "increasing BNP" (surface form) corresponds to "Heart failure" (normalized form). We defined cardiotoxicity using the following normalized forms: cardiac enlargement, cardiac dysfunction, cardiomyopathy, and heart failure. Figure 5 illustrates the NER and EN processes using MedNERN, while Supplementary Table S29 shows the cardiotoxicity-related portion of the normalization dictionary. Cardiotoxicity occurrence was determined when a named entity of the disease name class with a positive factuality attribute matched any surface form in Supplementary Table S29. The document recording date was considered the occurrence date. The performance of MedNERN in extracting positive disease name classes (macro-F value) is reported as 59.21% for case reports and 84.88% for radiology reports[56].

Covariates

We considered patient demographics (age: <65 years, 65–79 years, and ≥80 years; and sex), eight cancer sites (digestive organs, respiratory and intrathoracic organs, bones and articular cartilage, mesothelial and soft tissue, breast, female genital organs, urinary tract, lymphoid, hematopoietic and related tissues) and metastatic defined by ICD-10, and three comorbidities (diabetes, ischemic heart disease, and heart failure) as covariate candidates. Ischemic heart disease and heart failure were defined by ICD-10 codes, while diabetes was defined by a combination of ICD-10 and ATC codes. Additionally, the use of six cardiotoxic anticancer drugs other than AC (nitrogen mustard analogs, pyrimidine analogs, vinca alkaloids and analogs, etoposide, taxanes, and platinum compounds), opioid, diuretics (excluding AAs), and five cardioprotective drugs (ARB/ACEIs, BBs, statins, AAs, and CCBs) was included. For each comparative experiment, covariates with >5% prevalence in the exposure group were used for PSM stabilization. However, cardioprotective drug use was considered an important covariate and included in PSM even if its prevalence was <5%. Patient demographics, cancer sites, and comorbidities were extracted from the latest DPC data prior to the first AC administration. The concomitant medication period was defined as 60 days before the first AC administration, similar to that in the exposure group. The definitions of each disease based on ICD-10 codes and ATC classification codes are provided in Supplementary Table S30.

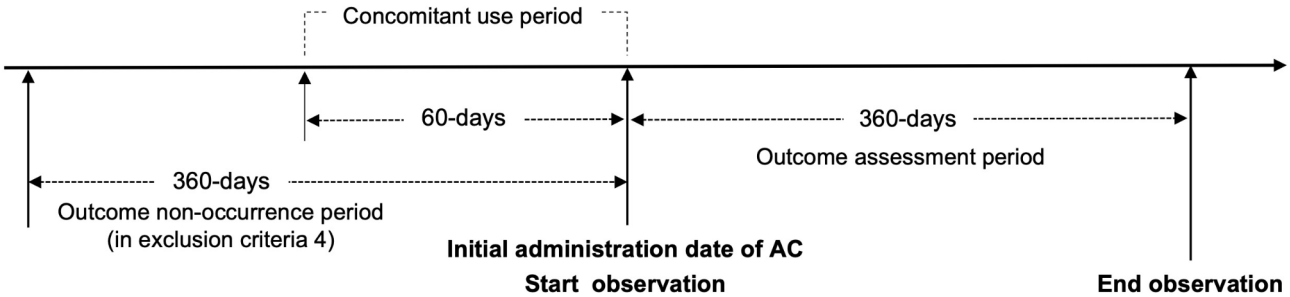

Fig. 4. Timeline of combination use and observation periods. The period of concomitant medication use was defined as 60 days prior to the initial AC administration date. The observation start date was set as the initial AC administration date, and patients who experienced the outcome within 360 days prior to this date were excluded. The occurrence of outcomes was evaluated over the 360 days following the observation start date. *AC* anthracyclines.

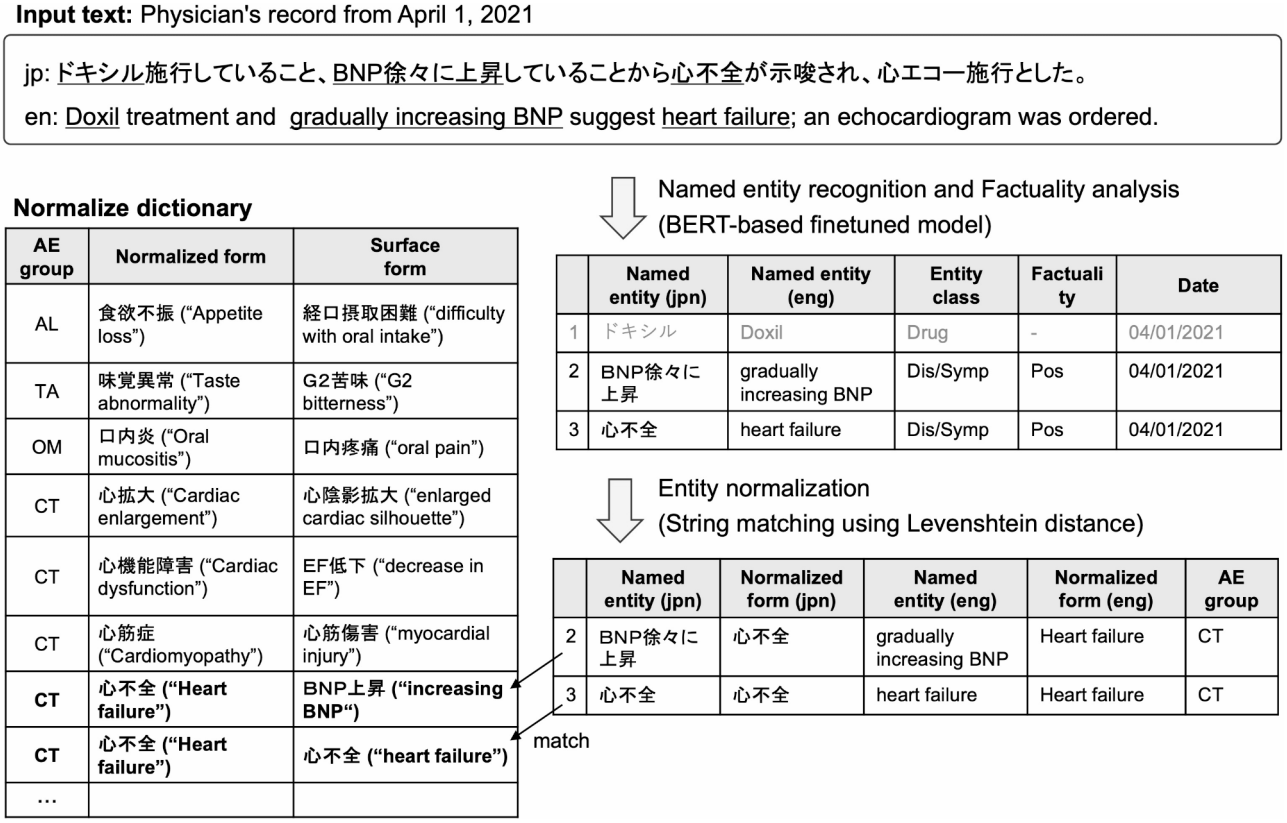

Fig. 5. Overview of named entity recognition (NER), factuality analysis (FA), and entity normalization (EN) processing. The process begins with NER and FA using a BERT-based fine-tuned model. This step extracts named entities from the input text and classifies the entity type (e.g., drug, disease/symptom) and factuality type. In this example, “ドキシル” (Doxil) is identified as a drug, “BNP徐々に上昇” (gradually increasing BNP) as a positive symptom, and “心不全” (heart failure) as a positive symptom. Next, the EN step utilizes string matching with the Levenshtein distance to align the named entities with normalized terms from the dictionary. For instance, “BNP徐々に上昇” (gradually increasing BNP) is matched with “BNP上昇” (increasing BNP) in the dictionary and normalized to “Heart failure,” which falls under the AE group “CT” (Cardiotoxicity). Similarly, “心不全” (heart failure) is matched with “心不全” (heart failure) in the dictionary and normalized to “Heart failure.” In the figure, “jpn” refers to Japanese and “eng” refers to English. The character string corresponding to “eng” is for explanatory purposes and does not appear in the actual analysis. *AE* adverse event, *AL* appetite loss, *BERT* bidirectional encoder representations from transformers, *CT* cardiotoxicity, *OM* oral mucositis, *TA* taste abnormality.

Statistical analysis

The effects of ARB/ACEI, BB, statin, and CCB on AIC were compared using PSM. There were no covariates with missing values. Propensity scores (PSs) for ARB/ACEI, BB, statin, and CCB prescription were estimated using the covariate balancing PS, which simultaneously optimizes PS estimation and covariate balance, thereby improving covariate balance more efficiently than traditional logistic regression-based methods[57,58]. Using the estimated PS, a control group corresponding to the exposure group was identified through 1:1 nearest-neighbor matching without replacement. Baseline characteristics between the exposure and control groups were compared using ASD before and after PSM. ASD < 10% was considered indicative of negligible imbalance between the baseline characteristics of the exposure and control groups. In such cases, a univariate Cox PH model was used to model the time to cardiotoxicity occurrence, estimating HRs and 95% CIs. If any baseline characteristic exhibited an ASD of > 10%, it was adjusted for in a multivariate Cox PH model to estimate HRs[59,60]. Proportional hazards assumptions were verified by visually inspecting log-transformed cumulative incidence curves. The p-values were two-sided, with a p-value of < 0.05 considered statistically significant. To assess the adequacy of sample sizes for each analysis, we conducted a post hoc power analysis by evaluating the adjusted models. In this analysis, we set the significance level at $\alpha = 0.05$ and calculated the power based on the sample size, observed event rate during the observation period, and obtained HR. All statistical analyses were performed using Stata/MP version 18.0 (StataCorp, College Station, TX, USA).

Sensitivity analysis

Several sensitivity analyses were conducted. First, as AIC occurs dose-dependently, we estimated HRs according to the cumulative number of AC prescriptions and confirmed the proportional hazards assumption. The number

of prescriptions was used instead of the actual dosage because the prescription data used in this study did not reveal the actual administered AC dosages. Second, to verify the robustness with respect to differences in the concomitant medication period, we examined two scenarios: one where the concomitant period was extended to 90 days prior to the initial AC administration, and another where it was shortened to 30 days prior to the initial AC administration. In both scenarios, the definition of the concomitant medication period for baseline characteristics remained consistent. Third, as a negative control, we estimated the HRs for PPIs, which are not expected to have cardioprotective effects, using the same settings as the primary analysis.

Data availability

The datasets analyzed during the current study are not publicly available due to restrictions imposed by the research ethics committee of the Graduate School of Medicine and Faculty of Medicine, The University of Tokyo, as they contain sensitive patient information. In accordance with the policy of The University of Tokyo Hospital, disclosure of data is neither included in the ethics application nor permitted in this study. For specific information requests, please contact the corresponding author.

Received: 28 August 2024; Accepted: 18 February 2025
Published online: 24 February 2025

References

1. Ashburn, T. T. & Thor, K. B. Drug repositioning: identifying and developing new uses for existing drugs. *Nat. Rev. Drug Discov.* **3**, 673–683 (2004).
2. Pushpakom, S. et al. Drug repurposing: progress, challenges and recommendations. *Nat. Rev. Drug Discov.* **18**, 41–58 (2019).
3. Xu, H., Li, J., Jiang, X. & Chen, Q. Electronic health records for drug repurposing: current status, challenges, and future directions. *Clin. Pharmacol. Ther.* **107**, 712–714 (2020).
4. Hurle, M. R. et al. Computational drug repositioning: from data to therapeutics. *Clin. Pharmacol. Ther.* **93**, 335–341 (2013).
5. Xu, H. et al. Validating drug repurposing signals using electronic health records: a case study of Metformin associated with reduced cancer mortality. *J. Am. Med. Inf. Assoc.* **22**, 179–191 (2015).
6. Kraus, A. L. et al. Real-world data of Palbociclib in combination with endocrine therapy for the treatment of metastatic breast cancer in men. *Clin. Pharmacol. Ther.* **111**, 302–309 (2022).
7. Shuey, M. M. et al. A genetically supported drug repurposing pipeline for diabetes treatment using electronic health records. *EBioMedicine* **94**, 104674 (2023).
8. Imai, S. et al. Using Japanese big data to investigate novel factors and their high-risk combinations that affect vancomycin-induced nephrotoxicity. *Br. J. Clin. Pharmacol.* **88**, 3241–3255 (2022).
9. Zamami, Y. et al. Identification of prophylactic drugs for oxaliplatin-induced peripheral neuropathy using big data. *Biomed. Pharmacother.* **148**, 112744 (2022).
10. Nadkarni, P. M. Drug safety surveillance using de-identified EMR and claims data: issues and challenges. *J. Am. Med. Inf. Assoc.* **17**, 671–674 (2010).
11. Classen, D. C. et al. Global trigger tool' shows that adverse events in hospitals May be ten times greater than previously measured. *Health Aff. (Millwood).* **30**, 581–589 (2011).
12. Perez-Lopez, R., Reis-Filho, J. S. & Kather, J. N. A framework for artificial intelligence in cancer research and precision oncology. *NPJ Precis. Oncol.* **7**, 43 (2023).
13. Khurshid, S. et al. Cohort design and natural Language processing to reduce bias in electronic health records research. *NPJ Digit. Med.* **5**, 47 (2022).
14. Guevara, M. et al. Large Language models to identify social determinants of health in electronic health records. *NPJ Digit. Med.* **7**, 6 (2024).
15. Vaswani, A. et al. Attention is all you need. *ArXiv* (2017).
16. Devlin, J., Chang, M., Lee, K. & Toutanova, K. Pre-training of deep bidirectional Transformers for Language Understanding. *ArXiv*. https://doi.org/10.48550/arXiv.1810.04805 (2018).
17. Lee, J. et al. BioBERT: a pre-trained biomedical Language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2020).
18. Kawazoe, Y., Shibata, D., Shinohara, E., Aramaki, E. & Ohe, K. A clinical specific BERT developed using a huge Japanese clinical text corpus. *PLoS One*. **16**, e0259763 (2021).
19. Kim, Y. et al. A pre-trained BERT for Korean medical natural Language processing. *Sci. Rep.* **12**, 13847 (2022).
20. Zhou, S., Kang, H., Yao, B. & Gong, Y. An automated pipeline for analyzing medication event reports in clinical settings. *BMC Med. Inf. Decis. Mak.* **18**, 113 (2018).
21. El-allaly, E., Sarrouti, M. & En-Nahnahi, N. Ouatik El Alaoui S. MTTLADE: A multi-task transfer learning-based method for adverse drug events extraction. *Inf. Process. Manag.* **58**, 102473 (2021).
22. Mahendran, D. & McInnes, B. T Extracting adverse drug events from clinical notes. *AMIA Jt. Summits Transl. Sci. Proc.* **2021**, 420–429 (2021).
23. Wu, H. et al. Chinese-named entity recognition from adverse drug event records: radical embedding-combined dynamic embedding-based BERT in a bidirectional long short-term conditional random field (Bi-LSTM-CRF) model. *JMIR Med. Inf.* **9**, e26407 (2021).
24. Narayanan, S. et al. A contextual multi-task neural approach to medication and adverse events identification from clinical text. *J. Biomed. Inf.* **125**, 103960 (2022).
25. Kim, S. et al. Automatic extraction of comprehensive drug safety information from adverse drug event narratives in the Korea adverse event reporting system using natural Language processing techniques. *Drug Saf.* **46**, 781–795 (2023).
26. Booth, L. K., Redgrave, R. E., Folaranmi, O., Gill, J. H. & Richardson, G. D. Anthracycline-induced cardiotoxicity and senescence. *Front. Aging.* **3**, 1058435 (2022).
27. Han, X., Zhou, Y. & Liu, W. Precision cardio-oncology: Understanding the cardiotoxicity of cancer therapy. *NPJ Precis. Oncol.* **1**, 31 (2017).
28. Gao, Y. et al. ACEI/ARB and beta-blocker therapies for preventing cardiotoxicity of antineoplastic agents in breast cancer: a systematic review and meta-analysis. *Heart Fail. Rev.* **28**, 1405–1415 (2023).
29. Lewinter, C. et al. A systematic review and meta-analysis of beta-blockers and renin-angiotensin system inhibitors for preventing left ventricular dysfunction due to anthracyclines or trastuzumab in patients with breast cancer. *Eur. Heart J.* **43**, 2562–2569 (2022).
30. Dong, H. et al. Can ACEI/ARB prevent the cardiotoxicity caused by chemotherapy in early-stage breast cancer? -a meta-analysis of randomized controlled trials. *Transl. Cancer Res.* **9**, 7034–7043 (2020).
31. Kalam, K. & Marwick, T. H. Role of cardioprotective therapy for prevention of cardiotoxicity with chemotherapy: a systematic review and meta-analysis. *Eur. J. Cancer.* **49**, 2900–2909 (2013).

32. Mir, A. et al. Efficacy and safety of cardioprotective drugs in chemotherapy-induced cardiotoxicity: an updated systematic review & network meta-analysis. *Cardiooncology* **9**, 10 (2023).
33. Ma, Y. et al. Beta-blockers for the primary prevention of anthracycline-induced cardiotoxicity: a meta-analysis of randomized controlled trials. *BMC Pharmacol. Toxicol.* **20**, 18 (2019).
34. He, D., Hu, J., Li, Y. & Zeng, X. Preventive use of beta-blockers for anthracycline-induced cardiotoxicity: A network meta-analysis. *Front. Cardiovasc. Med.* **9**, 968534 (2022).
35. Titus, A. et al. Statin for attenuating cardiotoxicity in patients receiving anthracyclines: a systematic review and meta-analysis. *Curr. Probl. Cardiol.* **48**, 101885 (2023).
36. Nabati, M., Janbabai, G., Esmailian, J. & Yazdani, J. Effect of Rosuvastatin in preventing chemotherapy-induced cardiotoxicity in women with breast cancer: a randomized, single-blind, placebo-controlled trial. *J. Cardiovasc. Pharmacol. Ther.* **24**, 233–241 (2019).
37. Neilan, T. G. et al. Atorvastatin for anthracycline-associated cardiac dysfunction: the STOP-CA randomized clinical trial. *JAMA* **330**, 528–536 (2023).
38. Thavendiranathan, P. et al. Statin to prevent early cardiac dysfunction in cancer patients at increased cardiotoxicity risk receiving anthracyclines. *Eur. Heart J. Cardiovasc. Pharmacother.* **9**, 515–525 (2023).
39. Hundley, W. G. et al. Statin and left ventricular ejection fraction following doxorubicin treatment. *NEJM Evid.* https://doi.org/10.1056/evidoa2200097 (2022).
40. Akpek, M. et al. Protective effects of spironolactone against anthracycline-induced cardiomyopathy. *Eur. J. Heart Fail.* **17**, 81–89 (2015).
41. Milei, J., Marantz, A., Alé, J., Vazquez, A. & Buceta, J. E. Prevention of adriamycin-induced cardiotoxicity by prenylamine: a pilot double blind study. *Cancer Drug Deliv.* **4**, 129–136 (1987).
42. Kraft, J. et al. Effects of verapamil on anthracycline-induced cardiomyopathy: preliminary results of a prospective multicenter trial. *Haematol. Blood Transfus.* **33**, 566–570 (1990).
43. Bansal, N. et al. Strategies to prevent anthracycline-induced cardiotoxicity in cancer survivors. *Cardiooncology* **5**, 18 (2019).
44. Ohishi, M. et al. Analysis of antihypertensive treatment using real-world Japanese data-the retrospective study of antihypertensives for Lowering blood pressure (REAL) study. *Hypertens. Res.* **42**, 1057–1067 (2019).
45. Ishida, T., Oh, A., Hiroi, S., Shimasaki, Y. & Tsuchihashi, T. Current prescription status of antihypertensive drugs in Japanese patients with hypertension: analysis by type of comorbidities. *Clin. Exp. Hypertens.* **41**, 203–210 (2019).
46. Monti, E., Paracchini, L., Piccinini, F., Rozza, A. & Villani, F. Effect of calcium inhibitors and calcium mobilizers on doxorubicin accumulation in rat myocardial tissue. *Pharmacol. Res. Commun.* **20**, 369–376 (1988).
47. Santostasi, G., Kutty, R. K. & Krishna, G. Increased toxicity of anthracycline antibiotics induced by calcium entry blockers in cultured cardiomyocytes. *Toxicol. Appl. Pharmacol.* **108**, 140–149 (1991).
48. Lan, Z. & Turchin, A. Impact of possible errors in natural Language processing-derived data on downstream epidemiologic analysis. *JAMIA Open.* **6**, ooad111 (2023).
49. Fusaroli, M. et al. The reporting of A disproportionality analysis for drug safety signal detection using individual case safety reports in pharmacovigilance (READUS-PV): explanation and elaboration. *Drug Saf.* **47**, 585–599 (2024).
50. Platt, R. et al. The new Sentinel Network–improving the evidence of medical-product safety. *N. Engl. J. Med.* **361**, 645–647 (2009).
51. Yamaguchi, M. et al. Establishment of the MID-NET® medical information database network as a reliable and valuable database for drug safety assessments in Japan. *Pharmacoepidemiol. Drug Saf.* **28**, 1395–1404 (2019).
52. Singhal, K. et al. Large Language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
53. Yasunaga, H., Ide, H., Imamura, T. & Ohe, K. Impact of the Japanese diagnosis procedure combination-based payment system on cardiovascular medicine-related costs. *Int. Heart J.* **46**, 855–866 (2005).
54. Yamana, H. et al. Validity of diagnoses, procedures, and laboratory data in Japanese administrative data. *J. Epidemiol.* **27**, 476–482 (2017).
55. Rosenbaum, P. R. & Rubin, D. B. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Am. Stat.* **39**, 33–38 (1985).
56. Yada, S., Nakamura, Y., Wakamiya, S. & Aramaki, E. Cross-lingual natural Language processing on limited annotated case/radiology reports in english and Japanese: insights from the Real-MedNLP workshop. *Methods Inf. Med.* (2024).
57. Imai, K. & Ratkovic, M. Covariate balancing propensity score. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **76**, 243–263 (2014).
58. Wyss, R. et al. The role of prediction modeling in propensity score estimation: an evaluation of logistic regression, bCART, and the covariate-balancing propensity score. *Am. J. Epidemiol.* **180**, 645–655 (2014).
59. Austin, P. C. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat. Med.* **28**, 3083–3107 (2009).
60. Nguyen, T. L. et al. Double-adjustment in propensity score matching analysis: choosing a threshold for considering residual imbalance. *BMC Med. Res. Methodol.* **17**, 78 (2017).

Acknowledgements

This study was funded by JST CREST [grant number JPMJCR22N1] and JSPS KAKENHI [grant number 23H03492], and partially funded by the Progress of the Next Cross-ministerial Strategic Innovation Promotion Program (SIP) on "Integrated Health Care System" [grant number JPJ012425]. The funder played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

Author contributions

YK: Conceptualization, Methodology, Resources, Writing – original draft, Writing – review and editing, Funding acquisition. MT: Conceptualization, Investigation, Writing – review and editing. KS: Software, Methodology, Data Curation, Writing – original draft. TS: Methodology, Formal analysis. ES: Software. SY: Software. SW: Software. SI: Conceptualization, Investigation. EA Software, Funding acquisition, Supervision. SH: Conceptualization, Funding acquisition, Supervision. All authors read and approved the final manuscript.

Declarations

Competing interests

YK, KS, and ES belong to the Artificial Intelligence and Digital Twin in Healthcare, Graduate School of Medicine, University of Tokyo, which is an endowment department supported by an unrestricted grant from EM Systems, EPNextS, MRP Co., Ltd., SHIP HEALTHCARE HOLDINGS, Inc., SoftBank Corp., and NEC Corporation; these organizations had no control over the interpretation, writing, or publication of this work. All authors declare no financial or non-financial competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-91187-6.

Correspondence and requests for materials should be addressed to Y.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.