Received: 18 January 2023 | Revised: 16 April 2023 | Accepted: 27 April 2023

DOI: 10.1002/ijc.34606

CANCER EPIDEMIOLOGY

IJC INTERNATIONAL JOURNAL of CANCER

Large-scale real-world data analyses of cancer risks among patients with rheumatoid arthritis

Feicheng Wang[1] | Nathan Palmer[2] | Kathe Fox[2] | Katherine P. Liao[3] | Kun-Hsing Yu[2,4] | Samuel C. Kou[1,5]

[1]Department of Statistics, Harvard University, Cambridge, Massachusetts, USA

[2]Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, USA

[3]Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts, USA

[4]Department of Pathology, Brigham and Women's Hospital, Boston, Massachusetts, USA

[5]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA

Correspondence

Samuel C. Kou, Department of Statistics, Harvard University, Cambridge, MA 02138, USA.
Email: kou@stat.harvard.edu

Kun-Hsing Yu, Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA.
Email: kun-hsing_yu@hms.harvard.edu

Funding information

Blavatnik Center for Computational Biomedicine Award; Google Research Scholar Award; Harold and Duval Bowen Fund; National Institute of General Medical Sciences, Grant/Award Number: R35GM142879; National Institute of Arthritis and Musculoskeletal and Skin Diseases, Grant/Award Number: P30AR072577

Abstract

Rheumatoid arthritis (RA) affects 24.5 million people worldwide and has been associated with increased cancer risks. However, the extent to which the observed risks are related to the pathophysiology of rheumatoid arthritis or its treatments is unknown. Leveraging nationwide health insurance claims data with 85.97 million enrollees across 8 years, we identified 92 864 patients without cancers at the time of rheumatoid arthritis diagnoses. We matched 68 415 of these patients with participants without rheumatoid arthritis by sex, race, age and inferred health and economic status and compared their risks of developing all cancer types. By 12 months after the diagnosis of rheumatoid arthritis, rheumatoid arthritis patients were 1.21 (95% confidence interval [CI] [1.14, 1.29]) times more likely to develop any cancer compared with matched enrollees without rheumatoid arthritis. In particular, the risk of developing lymphoma is 2.08 (95% CI [1.67, 2.58]) times higher in the rheumatoid arthritis group, and the risk of developing lung cancer is 1.69 (95% CI [1.32, 2.13]) times higher. We further identified the five most commonly used drugs in treating rheumatoid arthritis, and the log-rank test showed none of them is implicated with a significantly increased cancer risk compared with rheumatoid arthritis patients without that specific drug. Our study suggested that the pathophysiology of rheumatoid arthritis, rather than its treatments, is implicated in the development of subsequent cancers. Our method is extensible to investigating the connections among drugs, diseases and comorbidities at scale.

KEYWORDS

bDMARDs, cancers, matching method, rheumatoid arthritis, TNF inhibitors

What's new?

Cancer risk is increased by chronic inflammation, a significant feature of rheumatoid arthritis (RA). While RA patients are at increased risk of cancer, however, the degree to which cancer risk can be attributed to RA pathophysiology or treatment remains uncertain. Here, the authors examined relationships between RA, RA treatments and risk of different cancer types. RA patients were 1.69 to 2.08 times more likely than those without RA to develop lymphoma or

Abbreviations: bDMARDs, biological disease-modifying antirheumatic drugs; CCI, Charlson comorbidity index; CI, confidence interval; CPT, current procedural terminology; EHR, electronic health record; HCPCS, Healthcare Common Procedure Coding System; ICD, International Statistical Classification of Diseases and Related Health Problems; NDC, national drug codes; PheWAS, phenome-wide association study; RA, rheumatoid arthritis; sDMARDs, synthetic disease-modifying antirheumatic drugs; SIR, standardized incidence ratio; TNF, tumor necrosis factor; TNFi, tumor necrosis factor inhibitors.

lung cancer within 1 year of RA diagnosis. No significant difference in risk was detected for other cancer types. Commonly used RA treatments were also unlikely to increase cancer risk.

1 | INTRODUCTION

Rheumatoid arthritis (RA) is the most common autoimmune inflammatory joint disease affecting 24.5 million people worldwide,[1] with an incidence of 25 to 50 new cases per 100 000 people per year.[2,3] Chronic inflammation is also a risk factor for malignancy. RA patients have approximately twice the average risk for developing lymphoma.[4] Chronic inflammatory stimulation of the immune system,[5] genetic predisposition,[6] and RA treatments that modulate immune responses[7] have been linked with increased lymphoma risk, and patients with poorly controlled RA have the highest risk of developing lymphoma.[8,9] However, the connections between RA and other types of cancers are less clear,[10] because it is difficult to conduct large-scale longitudinal studies to simultaneously investigate many long-term health outcomes among a group of RA patients. Due to the fact that RA elicits abnormal chronic inflammation and may affect immune surveillance of malignant cells,[11] immune dysregulation is expected to alter the risks of cancers outside of the immune system as well.

The observed associations between RA and cancers can arise from the direct autoimmune processes underpinning RA or from treatments that modulated the immune system. The role of RA drugs in the development of cancers, particularly lymphoma, has been a topic of heated debate.[8,12] By suppressing specific components of the immune system, some researchers hypothesized that biologics might increase cancer risk.[13,14] A cohort study showed that RA patients with biologic therapies had a standardized incidence ratio (SIR) of 2.9 for developing cancers, while the SIR among RA patients without biologics was 1.9.[15] Another early meta-analysis of nine randomized controlled trials reported that patients treated with TNFi are more likely to develop malignancy compared with the placebo group (pooled odds ratio = 3.3; 95% CI 1.2-9.1).[16] However, a recent systematic review found that patients on bDMARDs (biological disease-modifying antirheumatic drugs) did not have an increased risk of malignancies in general.[17] Studies in the UK and Australia found no difference in the risk of lymphoma for the TNFi vs the biological-naive group (HR 1.00; 95% CI 0.56 to 1.80)[7,18] or risk of solid cancers for the TNFi vs synthetic disease-modifying antirheumatic drugs (sDMARDs) (HR = 0.83; 95% CI 0.64 to 1.07).[19] Other studies observed that RA patients treated with anti-TNF antibody therapy experienced increased and dose-dependent risks of malignancies.[16,20] For non-TNFi bDMARDs, a multi-database study in the U.S. reported a slight increase in total malignancy risk associated with abatacept compared with other biologics.[21] In addition, a recent review article identified a need for further studies on the cancer risks related to RA treatments in order to guide patients and clinicians regarding the optimal choice of antirheumatic drugs.[22] Large-scale and systematic investigations are thus needed to further evaluate and quantify the potential adverse effects conferred by other common treatments for RA.

To address these gaps in knowledge, our study leverages population-level insurance claims data with 86 million participants from the U.S. to reassess the risk of developing all types of cancers in a contemporary prospective cohort. We systematically examined the relationship between RA and all cancer types defined by the International Statistical Classification of Diseases and Related Health Problems (ICD) codes,[23] and we conducted detailed analyses to quantify cancer risks attributable to RA treatments. Using the matching method in causal inference,[24] we examined the causal relationship between RA, common treatments of RA and the subsequent development of cancers. Due to the large amount of data we have, we were able to exactly match almost all available confounding variables. Compared with propensity score matching,[21] our approach is less sensitive to model assumptions and can effectively characterize the interactions between diseases and treatments and decipher the mechanisms underlying the observed clinical outcomes.

2 | MATERIALS AND METHODS

2.1 | Data source

We performed this study using de-identified member claims data from Aetna, containing 85.97 million unique member identifiers from North America, with insurance claims records from April 1, 2008, to December 31, 2019. The follow-up period of these members starts on the date of subscription to Aetna insurance and ends on the date of subscription cancelation. A patient can have noncontinuous follow-up periods if he or she subscribes to Aetna for several non-consecutive time periods. Patients who exit the Aetna insurance system are considered censored at the time of insurance discontinuation, because we are unable to track their clinical outcomes after that point. If these patients re-joined the insurance plan and were not selected into our study cohort in previous enrollment periods, they will be eligible for our study if their clinical profiles from the new enrollment time satisfy our inclusion and exclusion criteria.

The claims data include diagnostic codes encoded by the International Statistical Classification of Diseases and Related Health Problems ninth revision (ICD-9) and tenth revision (ICD-10) as well as the Current Procedural Terminology (CPT)[23-25] and Healthcare Common Procedure Coding System (HCPCS) treatment procedure codes of the patients for every service and procedure, together with the date of service. In addition, the claims dataset contains National Drug Codes (NDC) for the drugs prescribed and the date of dispensing. We further extracted participants' insurance enrollment status, age, sex, race (15% available) and zip codes from the dataset.

2.2 | RA identification

Phenome-wide association study (PheWAS) code can be used to identify the mapping between a specific disease and its corresponding list of ICD-9 and ICD-10 codes. Guided by the PheWAS codes and descriptions,[26] we curated a list of ICD-9 and ICD-10 codes to identify RA patients from the insurance claims dataset (see Table S1).

2.3 | Cancer identification

To classify cancer diagnostic codes into clinically relevant categories, we conducted a manual review of all cancer diagnostic codes in both ICD-9 and ICD-10. Specifically, we reviewed all codes in Chapter II of ICD-9 (neoplasms; codes range from 140 to 239) and all codes in Chapter II of ICD-10 (neoplasms; codes range from C00 to D48). We classified diagnostic codes in these chapters into 17 categories of cancers. Tables S2 and S3 listed the cancer types we identified.

2.4 | Censoring

To maintain the integrity of the medical records we used in our analyses, patients were censored at the time of the first discontinuation of their health insurance plans. Reasons for health insurance plan discontinuation include un-subscription and death. Any medical encounters or pharmacy records timestamped after the censoring date were excluded from our analysis.

2.5 | Inclusion and exclusion criteria for the study cohort

Figure 1 shows the flowchart of our cohort identification. We first identified 232 943 patients with at least three diagnosis codes of RA on different days within 18 consecutive months to reduce the impact of false positives in the diagnostic labels. Previous studies showed that using two or more ICD-10 codes of RA has a positive predictive value (PPV) of 69% to 82% and a sensitivity of 76% to 77% for identifying RA patients, and using three or more ICD-9 codes of RA has a PPV of 66%.[27-29] In addition, we excluded 30 804 patients who had cancer or autoimmune disease history before the first time RA was diagnosed to better identify the cancer risks associated with RA and the treatments of RA. Furthermore, we excluded 106 044 patients who have mentions of RA or cancer within 90 days of enrollment to remove patients who may have RA or cancer diagnoses before enrolling in Aetna insurance. Finally, we removed 3231 patients with incomplete baseline characteristics (eg, zip codes) and those with inconsistent data (eg, patients with RA diagnosis date after their last enrollment date). Our procedure ensures that the identified 92 864 patients have a high probability of developing RA during their health insurance coverage period and do not have cancer or other autoimmune diseases when RA is first diagnosed. We identified the un-exposed group (non-RA participants) via a similar protocol. We randomly subsampled 10% of the participants in our non-RA group to reduce the computation time required. The un-exposed group after the 10% subsampling still has 7 653 770 patients, which is more than 80 times larger compared with the exposed group.

2.6 | Matching methods

We first compared the cancer risks among RA and non-RA groups. In this analysis, we employed the matching methods from the causal inference literature[24] to balance the demographic and clinical factors of the two groups. The matching methods construct pair-wise matching between the two groups such that the matched RA and non-RA patients share similar features associated with baseline cancer risk. We then compute the ratio of the cancer incidence rate of RA and non-RA group to estimate the cancer risk difference attributable to RA. To increase the efficiency of our study, each patient in the RA group is matched to multiple patients in the non-RA group.[30,31] When applying the matching method, we matched exactly on sex, year of birth and race (when available). Patients of unknown races were matched with patients of unknown races. The non-RA patients' index date is set as the matched RA patient's first day of RA diagnosis. We also matched patients' general health status and healthcare utilization rate using 10-quantile bins of their average diagnosis counts per year and average hospital visits per year. We further inferred their income levels using the median income of the patients' zip codes and matched them with 10-quantile bins. These matching factors ensure that the matched patients have similar demographics, health conditions and socioeconomic status.[32,33] Thus, the observed differences between the RA and the non-RA groups could be attributed to RA or its downstream effects.

Since each RA patient can be matched with multiple non-RA participants, we reweighted each of the matched control samples by inverse probability weighting. For example, if five participants in the non-RA group were matched to the same patient in the RA group, we assign a weight of 1/5 to each of these five participants in the non-RA group. This one-to-many matching method maximizes the number of eligible participants in the non-RA group, thereby reducing the random variability of our analyses. In the end, 68 415 of 92 864 RA patients found a match, and 1 340 538 of 7 653 770 non-RA patients found a match.

2.7 | Identifying the effects of RA treatments on the overall risk of cancers

We designed a similar matching method to investigate the effects of RA treatments on the overall risk of cancers among RA patients. A total of 32 847 RA patients have medication prescription coverage in the insurance claims database. Given this limited sample size, we

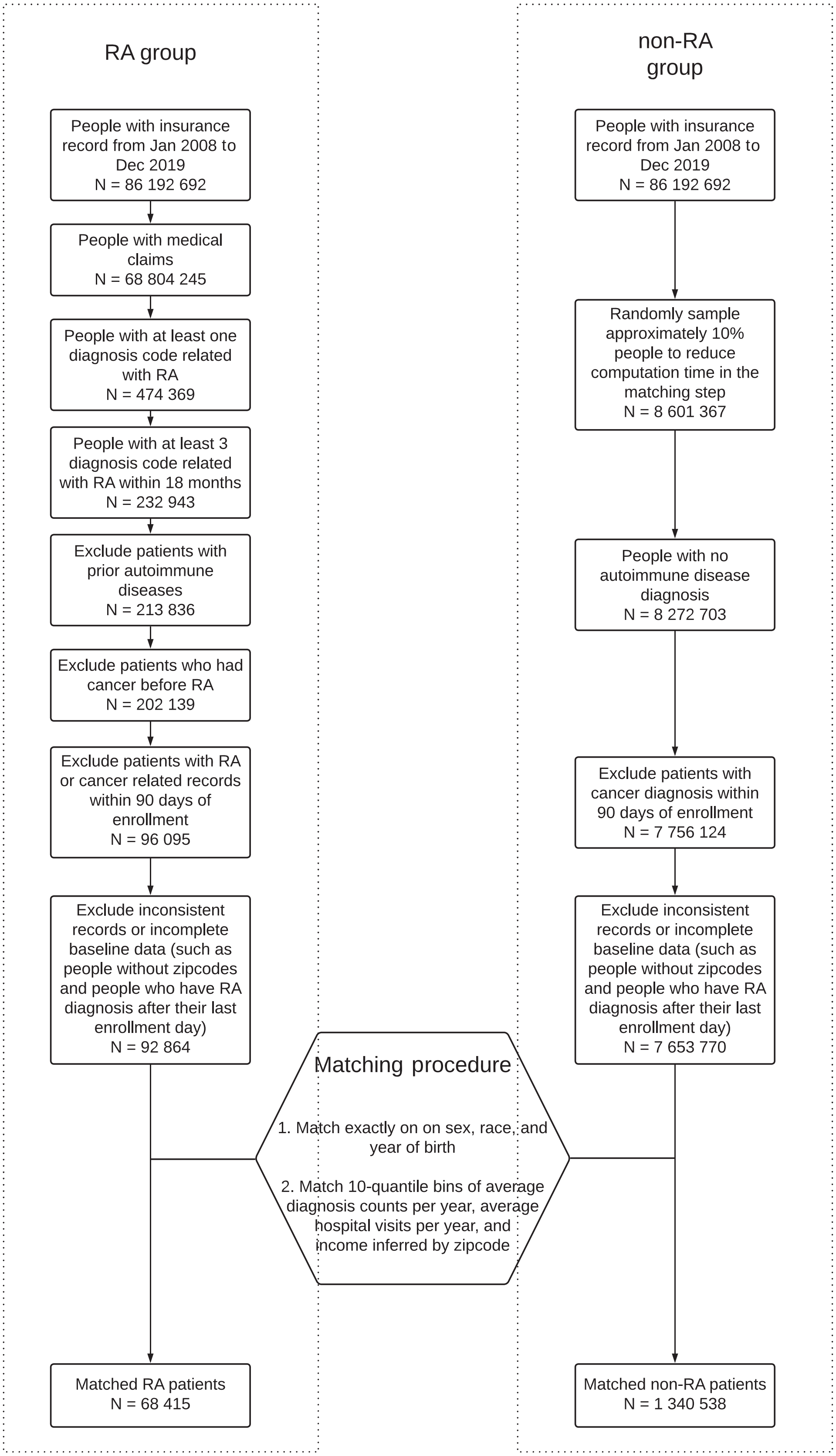

RA group
People with insurance record from Jan 2008 to Dec 2019
N = 86 192 692
People with medical claims
N = 68 804 245
People with at least one diagnosis code related with RA
N = 474 369
People with at least 3 diagnosis code related with RA within 18 months
N = 232 943
Exclude patients with prior autoimmune diseases
N = 213 836
Exclude patients who had cancer before RA
N = 202 139
Exclude patients with RA or cancer related records within 90 days of enrollment
N = 96 095
Exclude inconsistent records or incomplete baseline data (such as people without zipcodes and people who have RA diagnosis after their last enrollment day)
N = 92 864
Matched RA patients
N = 68 415
non-RA group
People with insurance record from Jan 2008 to Dec 2019
N = 86 192 692
Randomly sample approximately 10% people to reduce computation time in the matching step
N = 8 601 367
People with no autoimmune disease diagnosis
N = 8 272 703
Exclude patients with cancer diagnosis within 90 days of enrollment
N = 7 756 124
Exclude inconsistent records or incomplete baseline data (such as people without zipcodes and people who have RA diagnosis after their last enrollment day)
N = 7 653 770
Matched non-RA patients
N = 1 340 538
Matching procedure
1. Match exactly on on sex, race, and year of birth
2. Match 10-quantile bins of average diagnosis counts per year, average hospital visits per year, and income inferred by zipcode

FIGURE 1 A summary of the cohort derivation workflow for RA patients (exposed group) and matched non-RA patients (control group). Each inclusion, exclusion criteria and the number of patients after each step is shown.

focused on the five most commonly used medications for RA in this dataset: methotrexate (45.74% of RA patients received this drug), hydroxychloroquine (31.35%), TNFi (19.12%), leflunomide (10.06%) and sulfasalazine (8.77%). 70.35% of the 32 847 RA patients we studied used at least one of these five drugs. The baseline characteristics of these patients are summarized in Figure S4A.

TABLE 1 Patient characteristics

	Matched RA patients		Matched non-RA patients	
	N	%	N	%
All	68 415	100	1 340 538	100
Gender				
F	52 577	76.85	1 021 922	76.23
M	15 838	23.15	318 616	23.77
Age at RA diagnosis				
[0, 40]	11 417	16.69	230 246	17.18
[40, 50]	14 085	20.59	276 483	20.62
[50, 60]	19 060	27.86	374 602	27.94
[60, 70]	13 587	19.86	260 436	19.43
[70, 80]	6883	10.06	130 727	9.75
[80, 120]	3383	4.94	68 044	5.08
Zip-code inferred income				
[0, 35 000]	5223	7.63	101 904	7.6
[35 000, 50 000]	18 159	26.54	352 839	26.32
[50 000, 70 000]	22 778	33.29	446 458	33.3
[70 000, 100 000]	16 704	24.42	327 405	24.42
[100 000, 500 000]	5551	8.11	111 932	8.35
Frequency of receiving diagnostic codes before RA onset (per day)[a]				
[0.0, 0.03]	14 263	20.85	333 985	24.91
[0.03, 0.1]	18 677	27.3	349 583	26.08
[0.1, 0.3]	23 469	34.3	433 111	32.31
[0.3, 1.0]	10 691	15.63	198 679	14.82
[1.0, inf]	1315	1.92	25 180	1.88
Frequency of hospital visits before RA onset (per day)[b]				
[0.0, 0.01]	14 691	21.47	331 415	24.72
[0.01, 0.02]	11 073	16.19	207 342	15.47
[0.02, 0.05]	22 133	32.35	408 329	30.46
[0.05, 0.1]	13 102	19.15	247 388	18.45
[0.1, inf]	7416	10.84	146 064	10.9

Note: The index dates of non-RA participants are defined as their matched RA patients' date of RA diagnosis.

[a] A proxy of general health status.

[b] A proxy of healthcare utilization rate.

For each drug D, we identified RA patients treated with D after the initial diagnosis of RA as the exposed group and RA patients without RA-related drugs but with medication prescription coverage as the control group. The control group has a total of 15 499 patients. If a patient in the control group later received drug D, we censored this patient at the time he or she received drug D. This study design mitigates immortal time bias in electronic health record (EHR) analyses.[34] In the exposed group, we require drug D to be the first RA-related treatment the patient received. To better identify the effects of individual drugs, a patient is censored if he or she receives the second RA-related drug.

We matched the exposed and control groups based on their age at the first RA diagnosis, demographics (eg, sex, race, zip-code imputed median income levels) and clinical factors (general health conditions and healthcare utilization rate) mentioned above. Due to the more restricted sample size in this analysis among the RA patients, we developed a more relaxed matching scheme that used 5-quantile bins matching for continuous variables. Unmatched patients were not included in our analyses. To ensure that the baseline clinical characteristics of the exposed group and control groups are comparable, we examined the differences between the two groups. We compared the risk of developing cancers between the exposed and control groups using the log-rank test, and we corrected for multiple testing using the Benjamini-Hochberg procedure.[35]

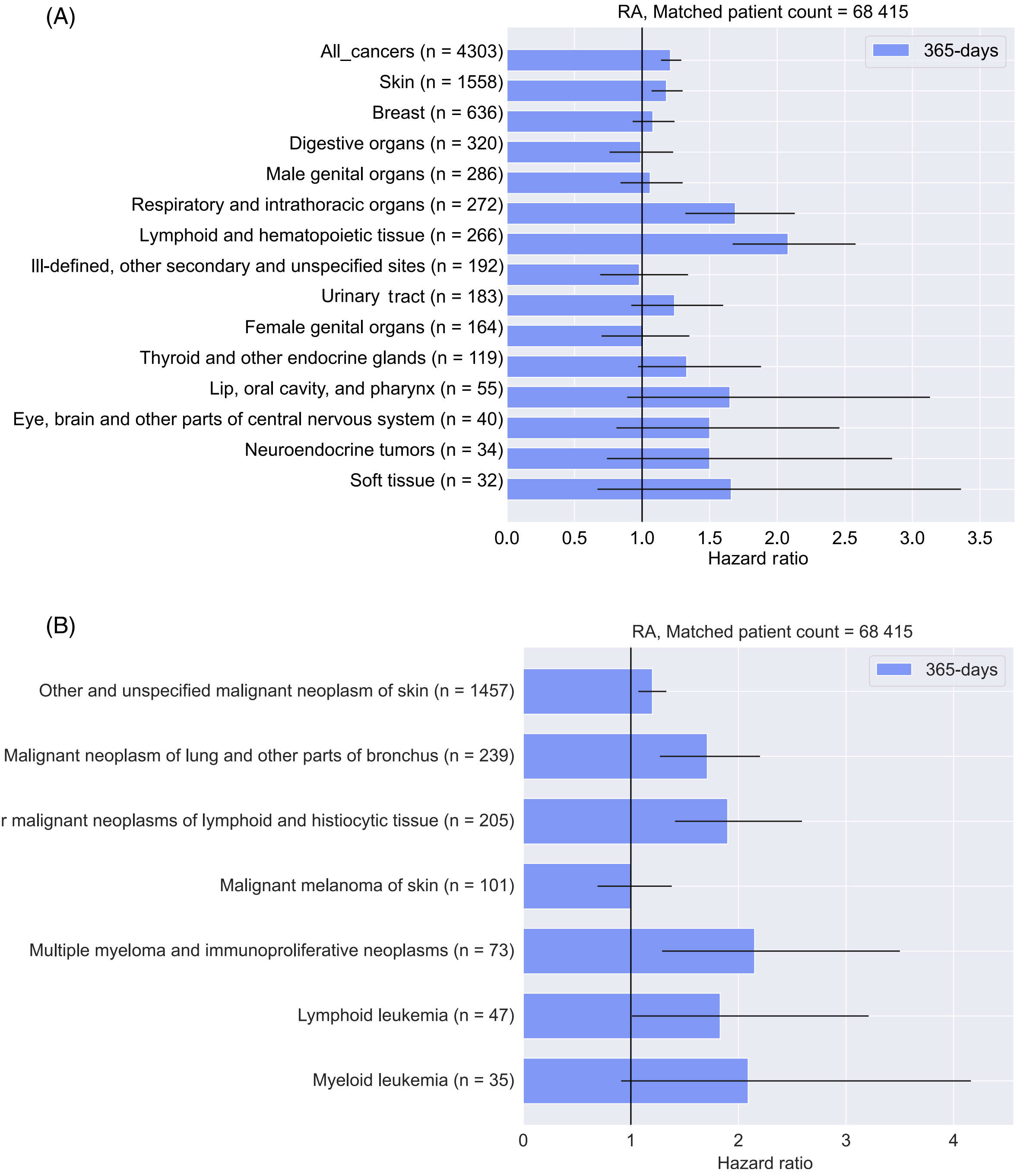


FIGURE 2 The hazard ratios of developing cancers at 365 days in RA and matched non-RA patients. (A) RA patients have increased risks of developing cancers overall, with particular enrichment in hematologic, lymphoid and respiratory cancers. (B) Cancer subtype analyses revealed that RA patients are more likely to develop skin, lung, lymphoid and histiocytic cancers, and multiple myeloma. The vertical line represents the hazard ratio of 1. The cancer types are ordered by their case counts. For example, skin cancer is more prevalent than breast cancer, so skin cancer appears at the top of the list. The horizontal segments represent the 2.5% to 97.5% confidence interval determined by 10 000 bootstrap[30] samples. [Color figure can be viewed at wileyonlinelibrary.com]

3 | RESULTS

3.1 | Overview of the study cohort and the trend of treatments

We identified 68 415 patients with RA from 85 972 617 participants in the insurance claims dataset. The average enrollment period for all participants is 3.8 years, which gives us an annual incidence of RA estimate of 21 per 100 000 people. Figure 1 shows the workflow for cohort identification. Detailed inclusion and exclusion criteria could be found in the Methods section. The average enrollment period after the patients' first RA diagnosis is approximately 3.5 years. Table 1 shows the characteristics of our study population, including gender, age, zip-code inferred income, the frequency of receiving ICD

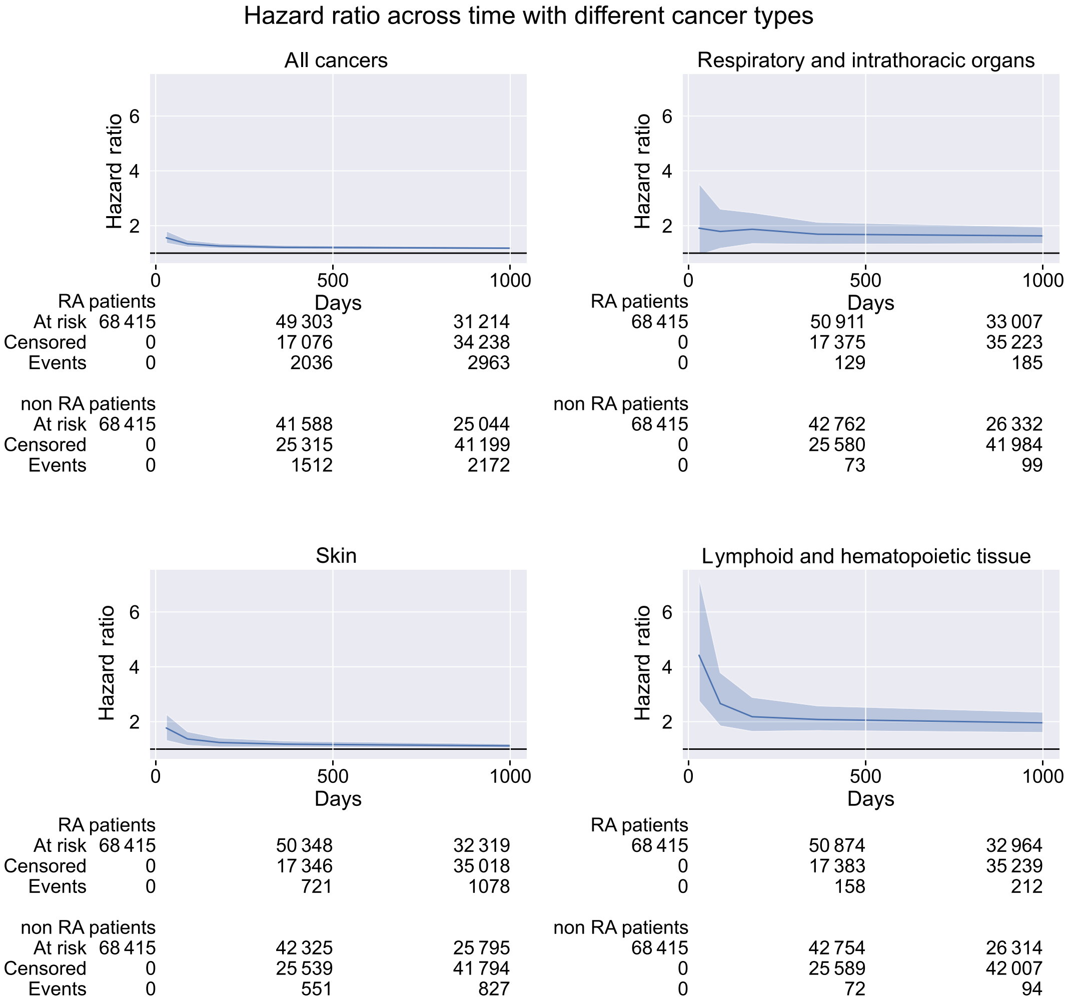


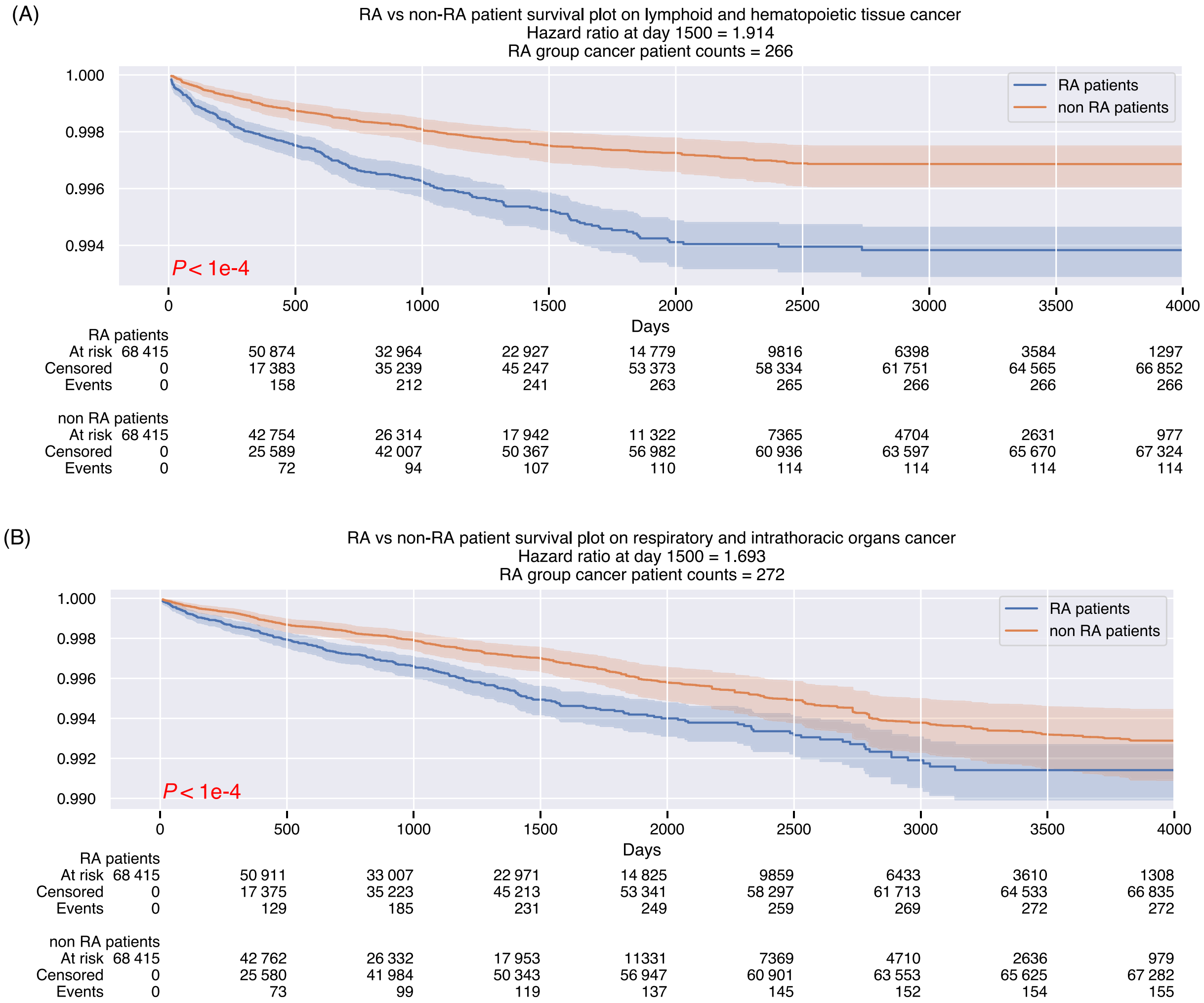
FIGURE 3 Hazard ratios of cancers after the diagnosis of RA. The hazard ratios are significantly >1 in the RA group compared with the matched control group throughout the time horizon we investigated. The horizontal black line represents a hazard ratio of 1. The solid curve is the estimated hazard ratio, and the shaded areas are 95% confidence intervals. [Color figure can be viewed at wileyonlinelibrary.com]

diagnostic codes (a proxy for general health status) and the frequency of hospital visits.

3.2 | RA and the risk of developing cancers

We first quantified the risk of developing cancers in the RA and non-RA groups. The overall probability of being diagnosed with cancers within 1 year is 2.57% for RA patients and 2.12% for non-RA patients, with a hazard ratio of 1.21 (95% confidence interval (CI) [1.14, 1.29]). We further computed the hazard ratio for each specific cancer type. A complete list of cancer types can be found in Tables S2 and S3. The cancer types with significantly increased risks among RA patients are cancers of the lymphoid and hematopoietic tissue (2.08; 95% CI [1.67, 2.58]), respiratory and intrathoracic organ cancer (1.69; 95% CI [1.32, 2.13]) and skin cancers (1.18; 95% CI [1.07, 1.3]). Many other cancer types, including cancers of the soft tissue and lip, oral cavity and pharynx cancers are also slightly enriched in the RA group; however, this enrichment did not achieve statistical significance (Figure 2A).

Beyond broad cancer categories such as skin and breast cancer, we conducted additional analyses to investigate the risk of developing specific types of cancer. To reduce the impact of multiple testing, our analyses focus on cancer categories with a hazard ratio significantly larger than 1 (ie, cancer categories whose horizontal confidence intervals in Figure 2A do not cross the vertical black line representing Hazard Ratio = 1). Three broad cancer categories met our significance criteria: skin cancer, lung cancer (respiratory and intrathoracic organs) and hematological cancer (lymphoid and hematopoietic tissue) (Figure 2B). Results showed that many cancer types under these phecode groups have a similar hazard ratio. For example, malignant neoplasm of the lung and other parts of the bronchus has a hazard ratio of 1.71 (95% CI [1.27, 2.20]) within a year after the RA diagnosis. Among the hematological cancers, other malignant neoplasms of lymphoid and histiocytic tissue possess a hazard ratio of 1.90 (95% CI [1.41, 2.59]), and multiple myeloma and immunoproliferative neoplasms have a hazard ratio of 2.15 (95% CI [1.29, 3.50]). Other cancer types in these phecode groups are observed in a smaller subset of RA patients, while the hazard ratio remains similar.



(A)

RA vs non-RA patient survival plot on lymphoid and hematopoietic tissue cancer
Hazard ratio at day 1500 = 1.914
RA group cancer patient counts = 266

	0	500	1000	1500	2000	2500	3000	3500	4000
RA patients									
At risk	68 415	50 874	32 964	22 927	14 779	9816	6398	3584	1297
Censored	0	17 383	35 239	45 247	53 373	58 334	61 751	64 565	66 852
Events	0	158	212	241	263	265	266	266	266
non RA patients									
At risk	68 415	42 754	26 314	17 942	11 322	7365	4704	2631	977
Censored	0	25 589	42 007	50 367	56 982	60 936	63 597	65 670	67 324
Events	0	72	94	107	110	114	114	114	114

(B)

RA vs non-RA patient survival plot on respiratory and intrathoracic organs cancer
Hazard ratio at day 1500 = 1.693
RA group cancer patient counts = 272

	0	500	1000	1500	2000	2500	3000	3500	4000
RA patients									
At risk	68 415	50 911	33 007	22 971	14 825	9859	6433	3610	1308
Censored	0	17 375	35 223	45 213	53 341	58 297	61 713	64 533	66 835
Events	0	129	185	231	249	259	269	272	272
non RA patients									
At risk	68 415	42 762	26 332	17 953	11331	7369	4710	2636	979
Censored	0	25 580	41 984	50 343	56 947	60 901	63 553	65 625	67 282
Events	0	73	99	119	137	145	152	154	155

FIGURE 4 Time to cancer onset in RA and matched non-RA patients. Day 0 is the day when the first diagnosis of RA is made or the matched date in non-RA patients. For each plot, the *P* value of the log-rank test is shown at the bottom left corner of the plot. The cancer type, the hazard ratio between RA and non-RA groups at 1500 days, and the total cancer patient count in the RA group are shown in the title of each figure panel. (A) Cancers of the lymphoid and hematopoietic tissue. (B) Cancers of the respiratory and intrathoracic organs. [Color figure can be viewed at wileyonlinelibrary.com]

We further investigate the hazard ratios across different time horizons in different cancer types. Figure 3 summarized the hazard ratios at 30, 90, 180, 365 and 1000 days from the first RA diagnosis across cancer types with the hazard ratio significantly larger than 1 on all days we examined. A similar plot summarizing all cancer types with at least 200 matched RA patients is shown in Figure S1. Tables S4 and S5 provide the 95% confidence intervals of the event (developing cancers) probability and hazard ratios at 365 days. In general, we observed a downward trend in the hazard ratios as time elapsed.

We plotted the time-to-event curves for the two cancer categories with the highest risk among RA patients: lymphoid and hematopoietic tissue cancer (Figure 4A) and respiratory and intrathoracic organ cancer (Figure 4B). Both have log-rank test *P* values <1e-4, showing that RA patients have a significantly higher risk of developing these cancers, compared with participants without RA. We also plotted the time-to-event curves for some cancer sub-categories that we studied in Figure 3 (Figure S2).

3.3 | Effects of RA drugs on the risk of developing cancers

We conducted a series of matched analyses to examine the effects of common RA drugs on the risks of cancers. The summary of the cohort derivation workflow is shown in Figure S3. We identified the five most commonly used drugs for RA treatment and conducted a log-rank test to evaluate the risk of developing cancers between the exposed and control groups. To ensure that the exposed and control groups are otherwise comparable, we examined the baseline characteristics of these two groups, and we found that the baseline confounders are very similar between the groups under comparison, except for the age distribution (Figure S4B). Our results showed that none of these drugs significantly increases the cancer risks among RA patients (Figure 5A, B; multiple testing was corrected using the Benjamini-Hochberg procedure[35]). We further conducted a series of analyses that

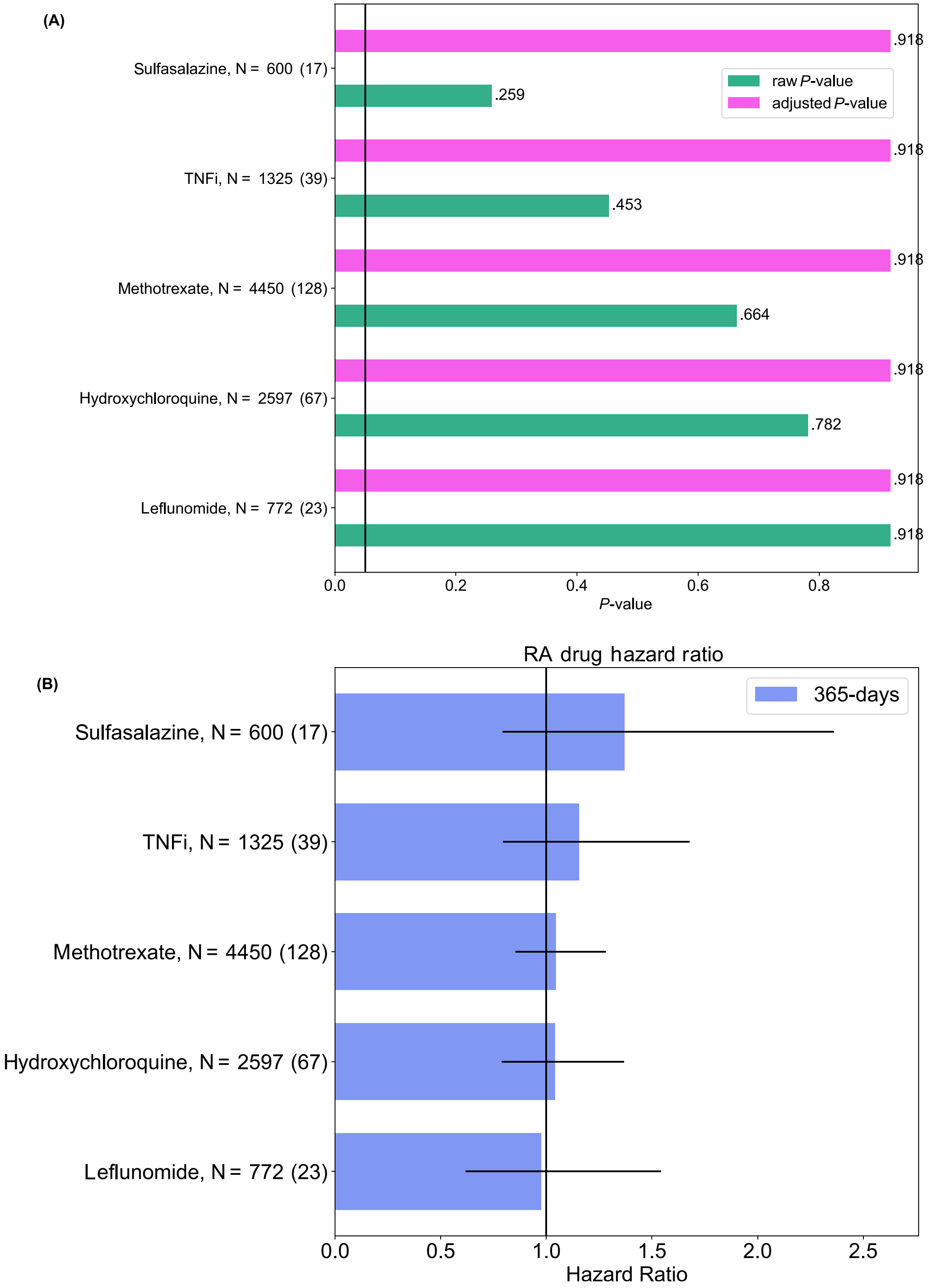


FIGURE 5 No significant associations were found between commonly used RA treatments and the development of any cancers among RA patients. (A) The *P* values and Benjamini-Hochberg-adjusted *P* values among the top five most commonly used drugs for RA treatment are shown. TNF inhibitors (TNFi's) include adalimumab, etanercept, certolizumab, golimumab and infliximab. The black vertical line represents *P* value = .05. For each RA treatment, we showed both the total number of patients with that treatment and the total number of patients in that group who later developed cancer. For example, "hydroxychloroquine, N = 2214(77)" indicates that there are 2214 patients who used hydroxychloroquine and among them, 77 are diagnosed with cancer. (B) The hazard ratios of developing cancers at 365 days between RA patients with RA drugs and matched RA patients without RA-related drugs. The horizontal segments represent the 2.5% to 97.5% confidence interval determined by 10 000 bootstrap[30] samples. [Color figure can be viewed at wileyonlinelibrary.com]

compares the effect (on cancer risks) difference among the five drugs of interest. We showed that there is no significant difference in cancer risks when comparing RA patients with any of the five drugs (drug D) with those without drug D (Figure S5A) or those without drug D but with other RA-related drugs (Figure S5B). We also showed that the baseline confounders are very similar among all five patient groups categorized by their drug type (Figure S4C).

4 | DISCUSSION

Our study leveraged concepts in causal inference to investigate the cancer risk attributable to the pathology of RA or RA treatments. One unique feature of our analyses is the use of large-scale observational datasets combined with matching methods to distill the contributions of diseases and subsequent treatments to the development of cancers. We employed health insurance records from 85.97 million unique members, which allows us to investigate and follow up with patients' evolving clinical phenotypes. We match many potential confounders, such as general health status (inferred by hospital visit and diagnosis rates) and household income status (inferred by zip codes). Our method is extensible to investigating the connections among drugs, diseases and comorbidities at scale.

We applied relatively stringent filters to ensure a low false identification rate in our patient group. We observed that the incidence of RA is 30 per 100 000 people in our study, which is slightly lower than the reported annual incidence of RA in the United States and northern European countries (approximately 40 per 100 000 people[36]). The minor difference in incidence rate may be partially due to the fact that participants of commercial health insurance (they are either employed or are dependents of employed people) are healthier than the general population. Compared with people without RA, we found that RA patients are 1.69 to 2.08 times more likely to develop lymphoma and lung cancers by 1 year after their first RA diagnosis. The risk of developing other types of cancers is not significantly different between the RA and the non-RA groups. To further distinguish the effects of RA drugs on the subsequent development of cancers and cancer predisposition among RA patients, we identified the five most commonly used drug groups for treating RA in our dataset: hydroxychloroquine, methotrexate, leflunomide, TNFi and sulfasalazine. None of these drugs significantly increased the cancer risk in RA patients compared with those who did not use these specific drugs. These results indicate that the observed increased risk for cancer may be driven more by immune dysregulation in RA, rather than RA therapies. Our findings provide reassuring information for both clinicians and patients when considering the risks and benefits of RA therapies.

Our studies expand the prior literature that shows the association between RA and risk for cancer. Prior studies have demonstrated that patients with RA appear to have a higher risk of lymphoma, lung cancer and skin cancer and a potentially decreased risk for colorectal and breast cancer compared with the general population.[20,37-40] Our analyses leverage a large electronic health record database and systematically investigate the contributions of RA pathology and subsequent treatment to the development of cancers. Our findings agree with previous findings on the increased lymphoma risk.[8,9] Additionally, we also found a significant risk increment in getting pulmonary and skin cancer. Interestingly, we found that RA patients are 1.08 times more likely to develop breast cancers within a year of their first RA diagnosis, although RA patients have a lower breast cancer risk (hazard ratio 0.87) over the entire time horizon we observed (4000 days), which is consistent with previous studies.[41] This observation may stem from the fact that patients just diagnosed with RA may become more health aware and more likely to discover existing slow-growing cancers. This pattern of decreasing hazard ratio over time is consistent across cancer types (Figures 3 and S1).

Several previous studies attempted to investigate the association between RA treatments and cancer. As an illustration, an early case series study using the MedWatch post-market adverse event surveillance system identified 26 patients who were treated with anti-TNF therapy and later developed lymphoproliferative disorders.[42] Follow-up studies and a meta-analysis reported a dose-dependent risk of malignancies among RA patients treated with anti-TNF therapy.[16] However, many recent studies showed no significant difference in lymphoma risk between RA patients who received anti-TNF and those treated with other drugs.[15,43-46] In addition, another study suggested that biologic therapy of RA is associated with increased risk for skin cancers, but not for solid tumors or lymphoproliferative malignancies.[47] One recent study found no skin cancer risk difference between RA patients initiating methotrexate vs those receiving hydroxychloroquine.[48] Most of these prior observational studies only adjusted for age and sex.[15,43,44] Other potential confounders could contribute to the observed differences. Our analyses systematically compare RA patients treated with common drugs with the matched patients receiving other forms of treatments. Using concepts in causal inference and the large sample size from the nationwide health insurance claims dataset, we showed that these drugs did not contribute significantly to the observed cancer risk. Thus, the increased cancer incidence among RA patients likely stems from the immune system disruptions related to RA. Controlling the disease activity and severity of RA through clinical follow-up and treatment may facilitate cancer risk mitigation in RA patients.

Our analyses also revealed a previously unreported downward trend in the hazard ratios among the cancer types associated with RA. There are several potential explanations for this trend. First, the impact of RA may diminish over time if the disease is under treatment, leading to lower cancer risk. The second possible explanation is that RA patients with high susceptibility to cancers also have a higher risk of dying from other causes or switching health insurance plans. In addition, diminishing statistical power on longer duration of follow-up could also explain a part of this trend. Future studies can further investigate the long-term cancer risks using registry data.

4.1 | Limitations

One limitation of our study is the limited granularity of ICD codes. ICD codes only provide a crude disease description but not the severity or the anatomical involvement of the diseases. For example, it is difficult to quantify the cancer risk difference between patients with severe RA and those with mild RA using health insurance claims data. In addition, our requirement of three separate RA diagnostic codes within 18 months decreases the false identification rate of RA but introduces a potential caveat in selecting patients with severe RA. Furthermore, we are unable to match the Charlson Comorbidity Index (CCI), an established mortality predictor using concurrent medical conditions, due to the constraints of our data use agreement. Lastly, many known risk factors for cancers, such as smoking, chronic alcoholism, red meat intake and exposure to asbestos, are not reliably documented in the insurance claims datasets. Future studies that use data from electronic medical notes can increase the granularity of the disease description. Because insurance claims datasets and electronic health records only cover a period of time in a patient's life span, they will have limited ability in long-term risk assessment. Although we only focused on the five most commonly used RA drugs due to statistical power considerations, future research can investigate the effects of other medications used in RA patients.

5 | CONCLUSIONS

Our study demonstrated that the pathophysiological changes related to RA, rather than RA therapy, are likely the main contributor to the increased risk of cancers observed in RA patients. Using the population-level insurance claims dataset, we found that the risk of lung cancer and lymphoma is the highest among patients with RA. Our approaches demonstrated the power of using nationwide EHRs in identifying the potential factors leading to major health outcomes. Future studies can employ similar methods to monitor the risk of developing cancers among other high-risk patient populations.

AUTHOR CONTRIBUTIONS

Feicheng Wang: conceptualization, data curation, formal analysis, investigation, methodology, validation, visualization, writing-original draft, writing-review & editing. Nathan Palmer: data curation, resources, software. Kathe Fox: project administration, resources. Katherine P. Liao: conceptualization, methodology, writing—original draft. Kun-Hsing Yu: conceptualization, funding acquisition, methodology, project administration, resources, supervision, writing-original draft, writing-review & editing. Samuel C. Kou: conceptualization, funding acquisition, methodology, project administration, resources, supervision, writing-review & editing. The work reported in the paper has been performed by the authors, unless clearly specified in the text.

FUNDING INFORMATION

Kun-Hsing Yu is supported in part by the National Institute of General Medical Sciences (NIGMS) grant R35GM142879, the Google Research Scholar Award and the Blavatnik Center for Computational Biomedicine Award. Katherine P. Liao is supported by the Harold and Duval Bowen Fund and the National Institute of Arthritis and Musculoskeletal and Skin Diseases P30 AR072577. The funders had no role in the design and conduct of the study; collection, management, analysis and interpretation of the data; preparation, review or approval of the manuscript; and decision to submit the manuscript for publication.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding authors upon reasonable request.

ETHICS STATEMENT

Written informed consent was obtained for all de-identified population-level studies from all participants at the time of enrollment in the health insurance plan. The informed consent was not specific to the study. Our study that involves secondary use of the de-identified health insurance data is approved by the Harvard Medical School Institutional Review Board (IRB20-0957).

ORCID

Kun-Hsing Yu https://orcid.org/0000-0001-9892-8218

REFERENCES

1. GBD 2015 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet*. 2016;388:1545-1602.
2. Alamanos Y, Voulgari PV, Drosos AA. Incidence and prevalence of rheumatoid arthritis, based on the 1987 American College of Rheumatology criteria: a systematic review. *Semin Arthritis Rheum*. 2006 Dec;36(3):182-188.
3. Uhlig T, Moe RH, Kvien TK. The burden of disease in rheumatoid arthritis. *PharmacoEconomics*. 2014;32:841-851. doi:10.1007/s40273-014-0174-6
4. Mercer LK, Regierer AC, Mariette X, et al. Spectrum of lymphomas across different drug treatment groups in rheumatoid arthritis: a European registries collaborative project. *Ann Rheum Dis*. 2017;76(12):2025-2030.
5. Baecklund E, Iliadou A, Askling J, et al. Association of chronic inflammation, not its treatment, with increased lymphoma risk in rheumatoid arthritis. *Arthritis Rheum*. 2006;54(3):692-701.
6. Nieters A, Beckmann L, Deeg E, Becker N. Gene polymorphisms in Toll-like receptors, interleukin-10, and interleukin-10 receptor alpha and lymphoma risk. *Genes Immun*. 2006 Dec;7(8):615-624.
7. Mercer LK, Galloway JB, Lunt M, et al. Risk of lymphoma in patients exposed to antitumour necrosis factor therapy: results from the British Society for rheumatology biologics register for rheumatoid arthritis. *Ann Rheum Dis*. 2017;76(3):497-503.
8. Klein A, Polliack A, Gafter-Gvili A. Rheumatoid arthritis and lymphoma: incidence, pathogenesis, biology, and outcome. *Hematol Oncol*. 2018;36(5):733-739.
9. Franks AL, Slansky JE. Multiple associations between a broad spectrum of autoimmune diseases, chronic inflammatory diseases and cancer. *Anticancer Res*. 2012 Apr;32(4):1119-1136.
10. Bhandari B, Basyal B, Sarao MS, Nookala V, Thein Y. Prevalence of cancer in rheumatoid arthritis: epidemiological study based on the National Health and nutrition examination survey (NHANES). *Cureus*. 2020;12(4):e7870.
11. Swann JB, Smyth MJ. Immune surveillance of tumors. *J Clin Invest*. 2007;117:1137-1146. doi:10.1172/jci31405
12. Wadström H, Frisell T, Askling J, Anti-Rheumatic Therapy in Sweden (ARTIS) Study Group. Malignant neoplasms in patients with rheumatoid arthritis treated with tumor necrosis factor inhibitors, tocilizumab, abatacept, or rituximab in clinical practice: a Nationwide cohort study from Sweden. *JAMA Intern Med*. 2017;177(11):1605-1612.
13. Solomon DH, Kremer JM, Fisher M, et al. Comparative cancer risk associated with methotrexate, other non-biologic and biologic

disease-modifying anti-rheumatic drugs. *Semin Arthritis Rheum*. 2014; 43(4):489-497.

14. Piovani D, Danese S, Peyrin-Biroulet L, Nikolopoulos GK, Bonovas S. Systematic review with meta-analysis: biologics and risk of infection or cancer in elderly patients with inflammatory bowel disease. *Aliment Pharmacol Ther*. 2020;51:820-830. doi:10.1111/apt.15692
15. Wolfe F, Michaud K. Lymphoma in rheumatoid arthritis: the effect of methotrexate and anti-tumor necrosis factor therapy in 18 572 patients. *Arthritis Rheum*. 2004;50(6):1740-1751.
16. Bongartz T, Sutton AJ, Sweeting MJ, Buchan I, Matteson EL, Montori V. Anti-TNF antibody therapy in rheumatoid arthritis and the risk of serious infections and malignancies: systematic review and meta-analysis of rare harmful effects in randomized controlled trials. *JAMA*. 2006;295(19):2275-2285.
17. Ramiro S, Sepriano A, Chatzidionysiou K, et al. Safety of synthetic and biological DMARDs: a systematic literature review informing the 2016 update of the EULAR recommendations for management of rheumatoid arthritis. *Ann Rheum Dis*. 2017;76(6):1101-1136.
18. Staples MP, March L, Hill C, Lassere M, Buchbinder R. Malignancy risk in Australian rheumatoid arthritis patients treated with anti-tumour necrosis factor therapy: an update from the Australian rheumatology association database (ARAD) prospective cohort study. *BMC Rheumatol*. 2019;3(1):1.
19. Mercer LK, Lunt M, Low ALS, et al. Risk of solid cancer in patients exposed to anti-tumour necrosis factor therapy: results from the British Society for rheumatology biologics register for rheumatoid arthritis. *Ann Rheum Dis*. 2015;74(6):1087-1093.
20. Geborek P. Tumour necrosis factor blockers do not increase overall tumour risk in patients with rheumatoid arthritis, but may be associated with an increased risk of lymphomas. *Ann Rheum Dis*. 2005;64: 699-703. doi:10.1136/ard.2004.030528
21. Simon TA, Boers M, Hochberg M, et al. Comparative risk of malignancies and infections in patients with rheumatoid arthritis initiating abatacept versus other biologics: a multi-database real-world study. *Arthritis Res Ther*. 2019;21(1):228.
22. Singh N, Li CI. Impact of rheumatoid arthritis and biologic and targeted synthetic disease modifying antirheumatic agents on cancer risk and recurrence. *Curr Opin Rheumatol*. 2021;33(3):292-299.
23. World Health Organization. *International Statistical Classification of Diseases and Related Health Problems: Alphabetical Index*. Geneve, Switzerland: World Health Organization; 2004.
24. Stuart EA. Matching methods for causal inference: a review and a look forward. *Stat Sci*. 2010;25(1):1-21.
25. Hirsch JA, Leslie-Mazwi TM, Nicola GN, et al. Current procedural terminology: a primer. *J Neurointerv Surg*. 2015;7(4):309-312.
26. Denny JC, Ritchie MD, Basford MA, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics*. 2010;26(9):1205-1210.
27. Ng B, Aslam F, Petersen NJ, Yu HJ, Suarez-Almazor ME. Identification of rheumatoid arthritis patients using an administrative database: a veterans affairs study. *Arthritis Care Res*. 2012;64(10):1490-1496.
28. Kim SY, Servi A, Polinski JM, et al. Validation of rheumatoid arthritis diagnoses in health care utilization data. *Arthritis Res Ther*. 2011;13(1):R32.
29. Curtis JR, Xie F, Zhou H, Salchert D, Yun H. Use of ICD-10 diagnosis codes to identify seropositive and seronegative rheumatoid arthritis when lab results are not available. *Arthritis Res Ther*. 2020;22(1):242.
30. Yu KH, Miron O, Palmer N, et al. Data-driven analyses revealed the comorbidity landscape of tuberous sclerosis complex. *Neurology*. 2018;91(21):974-976.
31. Yu KH, Palmer N, Fox K, et al. The phenotypical implications of immune dysregulation in fragile X syndrome. *Eur J Neurol*. 2020;27(3): 590-593.
32. Yang S, Yu KH, Palmer N, Fox K, Kou SC, Kohane IS. Autoimmune effects of lung cancer immunotherapy revealed by data-driven analysis on a Nationwide cohort. *Clin Pharmacol Ther*. 2019;107:388-396. doi:10.1002/cpt.1597
33. Wang F, Yang S, Palmer N, et al. Real-world data analyses unveiled the immune-related adverse effects of immune checkpoint inhibitors across cancer types. *NPJ Precis Oncol*. 2021;5(1):82.
34. Hernán MA, Sauer BC, Hernández-Díaz S, Platt R, Shrier I. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *J Clin Epidemiol*. 2016;79:70-75.
35. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Stat Soc*. 1995; 57:289-300. doi:10.1111/j.2517-6161.1995.tb02031.x
36. Myasoedova E, Crowson CS, Kremers HM, Therneau TM, Gabriel SE. Is the incidence of rheumatoid arthritis rising?: results from Olmsted County, Minnesota, 1955-2007. *Arthritis Rheum*. 2010;62:1576-1582. doi:10.1002/art.27425
37. Smitten AL, Simon TA, Hochberg MC, Suissa S. A meta-analysis of the incidence of malignancy in adult patients with rheumatoid arthritis. *Arthritis Res Ther*. 2008;10(2):R45.
38. Simon TA, Thompson A, Gandhi KK, Hochberg MC, Suissa S. Incidence of malignancy in adult patients with rheumatoid arthritis: a meta-analysis. *Arthritis Res Ther*. 2015;15(17):212.
39. Mellemkjaer L, Linet MS, Gridley G, Frisch M, Møller H, Olsen JH. Rheumatoid arthritis and cancer risk. *Eur J Cancer*. 1996;32A(10): 1753-1757.
40. Zintzaras E. The risk of lymphoma development in autoimmune diseases. *Archiv Intern Med*. 2005;165:2337. doi:10.1001/archinte.165.20.2337
41. Wadström H, Pettersson A, Smedby KE, Askling J. Risk of breast cancer before and after rheumatoid arthritis, and the impact of hormonal factors. *Ann rheum dis*. 2020;79(5):581-586.
42. Brown SL, Greene MH, Gershon SK, Edwards ET, Braun MM. Tumor necrosis factor antagonist therapy and lymphoma development: twenty-six cases reported to the Food and Drug Administration. *Arthritis Rheum*. 2002;46(12):3151-3158.
43. Askling J, Fored CM, Baecklund E, et al. Haematopoietic malignancies in rheumatoid arthritis: lymphoma risk and characteristics after exposure to tumour necrosis factor antagonists. *Ann Rheum Dis*. 2005; 64(10):1414-1420.
44. Askling J, Fored CM, Brandt L, et al. Risks of solid cancers in patients with rheumatoid arthritis and after treatment with tumour necrosis factor antagonists. *Ann Rheum Dis*. 2005;64(10):1421-1426.
45. Huss V, Bower H, Wadström H, Frisell T, Askling J, ARTIS Group. Short- and longer-term cancer risks with biologic and targeted synthetic disease-modifying antirheumatic drugs as used against rheumatoid arthritis in clinical practice. *Rheumatology*. 2022;61(5):1810-1818.
46. Hellgren K, Baecklund E, Backlin C, Sundstrom C, Smedby KE, Askling J. Rheumatoid arthritis and risk of malignant lymphoma: is the risk still increased? *Arthritis Rheumatol*. 2017;69(4):700-708.
47. Wolfe F, Michaud K. Biologic treatment of rheumatoid arthritis and the risk of malignancy: analyses from a large US observational study. *Arthritis Rheum*. 2007;56:2886-2895. doi:10.1002/art.22864
48. Solomon DH, Glynn RJ, Karlson EW, et al. Adverse effects of low-dose methotrexate: a randomized trial. *Ann Intern Med*. 2020;172(6):369-380.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Wang F, Palmer N, Fox K, Liao KP, Yu K-H, Kou SC. Large-scale real-world data analyses of cancer risks among patients with rheumatoid arthritis. *Int J Cancer*. 2023;153(6):1139-1150. doi:10.1002/ijc.34606