RMD Open
Rheumatic & Musculoskeletal Diseases

ORIGINAL RESEARCH

Prevalence and clinical characteristics of patients with rheumatoid arthritis with interstitial lung disease using unstructured healthcare data and machine learning

Jose A Román Ivorra,[1] Ernesto Trallero-Araguas,[2] Maria Lopez Lasanta,[2] Laura Cebrián,[3] Leticia Lojo,[3] Belén López-Muñíz,[4] Julia Fernández-Melon,[5] Belén Núñez,[6] Lucia Silva-Fernández,[5] Raúl Veiga Cabello,[7] Pilar Ahijado,[8] Isabel De la Morena Barrio,[9] Nerea Costas Torrijo,[10] Belén Safont,[11] Enrique Ornilla,[12] Juliana Restrepo,[13] Arantxa Campo,[14] Jose L Andreu,[15] Elvira Díez,[16] Alejandra López Robles,[17] Elena Bollo,[18] Diego Benavent,[19] David Vilanova,[20] Sara Luján Valdés,[21] Raul Castellanos-Moreira [21]

To cite:
RománIvorraJA,Trallero-AraguasE, Lopez Lasanta M, *et al*. Prevalence and clinical characteristics of patients with rheumatoid arthritis with interstitial lung disease using unstructured healthcare data and machine learning. *RMD Open* 2024;**10**:e003353. doi:10.1136/rmdopen-2023-003353

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/rmdopen-2023-003353).

Preliminary data of the RA-WILD study were presented as an oral presentation at EULAR Congress 2022.

Received 31 May 2023
Accepted 3 January 2024

For numbered affiliations see end of article.

Correspondence to
Dr Raul Castellanos-Moreira; raul.castellanos@bms.com

ABSTRACT

Objectives Real-world data regarding rheumatoid arthritis (RA) and its association with interstitial lung disease (ILD) is still scarce. This study aimed to estimate the prevalence of RA and ILD in patients with RA (RAILD) in Spain, and to compare clinical characteristics of patients with RA with and without ILD using natural language processing (NLP) on electronic health records (EHR).

Methods Observational case–control, retrospective and multicentre study based on the secondary use of unstructured clinical data from patients with adult RA and RAILD from nine hospitals between 2014 and 2019. NLP was used to extract unstructured clinical information from EHR and standardise it into a SNOMED-CT terminology. Prevalence of RA and RAILD were calculated, and a descriptive analysis was performed. Characteristics between patients with RAILD and RA patients without ILD (RAnonILD) were compared.

Results From a source population of 3 176 165 patients and 64 241 683 EHRs, 13 958 patients with RA were identified. Of those, 5.1% patients additionally had ILD (RAILD). The overall age-adjusted prevalence of RA and RAILD were 0.53% and 0.02%, respectively. The most common ILD subtype was usual interstitial pneumonia (29.3%). When comparing RAILD versus RAnonILD patients, RAILD patients were older and had more comorbidities, notably concerning infections (33.6% vs 16.5%, p<0.001), malignancies (15.9% vs 8.5%, p<0.001) and cardiovascular disease (25.8% vs 13.9%, p<0.001) than RAnonILD. RAILD patients also had higher inflammatory burden reflected in more pharmacological prescriptions and higher inflammatory parameters and presented a higher in-hospital mortality with a higher risk of death (HR 2.32; 95% CI 1.59 to 2.81, p<0.001).

Conclusions We found an estimated age-adjusted prevalence of RA and RAILD by analysing real-world data through NLP. RAILD patients were more vulnerable at the time of inclusion with higher comorbidity and inflammatory burden than RAnonILD, which correlated with higher mortality.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Rheumatoid arthritis (RA) is a prevalent form of inflammatory arthritis, affecting approximately 0.5% of the global population.

⇒ Interstitial lung disease (ILD) is a frequent and severe complication in patients with RA, but its prevalence and characterisation vary greatly depending on study design, population and method of ILD assessment.

INTRODUCTION

Rheumatoid arthritis (RA) is a systemic inflammatory disorder characterised by synovial inflammation and symmetric polyarthritis that leads to progressive joint erosion and eventual deformity.[1] It is the most common connective tissue disease and represents an increasing burden on global health resources.[2] Extra-articular manifestations of RA are common, affecting up to 40% of patients and interstitial lung disease (ILD) is one of the most frequent of them.[3–7]

The prevalence of ILD within RA has not yet been accurately estimated and varies significantly between studies, ranging from 2% to 40%.[8] Study design, diagnostic tools and methods of assessment of ILD, impact on the detection of the disease. While research using chest X-rays have estimated the prevalence of

WHAT THIS STUDY ADDS

⇒ This study is the first of its kind in rheumatology to use machine learning and natural language processing on unstructured data from electronic health records to estimate the prevalence and characterise the RA and RAILD (ILD in patients with RA) populations.

⇒ Among 13 958 patients with RA, 5.1% had RAILD; this yielded an estimated overall age-adjusted prevalence of RA and RAILD of 0.53% and 0.02%, respectively.

⇒ Patients with RA who develop ILD are older, with more comorbidities and higher inflammatory burden in comparison with patients with RA without ILD. They also show a higher in-hospital mortality.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ Understanding the peculiarities of patients with RA with ILD can aid healthcare providers in recognising and addressing the challenges faced by these patients.

⇒ This study highlights the strengths and usefulness of using real-world evidence in conducting epidemiological studies.

⇒ Natural language processing techniques in electronic health records may be particularly useful for analysing large volumes of real-world data and estimating infrequent events such as RAILD.

ILD at 5%, autopsy case series have found evidence of interstitial lung disease in 33% of patients with advanced RA.[9 10] Another study estimated ILD prevalence in RA at 41% by assessing the diffusion capacity of the lung for carbon monoxide.[11] Nevertheless, the advancement of high-resolution CT (HRCT) has led to the increasing use of this sensitive, non-invasive diagnostic tool, overcoming these observed disparities. Thus, Bongartz *et al* estimated the frequency of ILD among patients with RA between 4.0% and 7.9% using HRCT.[3] However, symptomatic ILD has been described in approximately 10% of patients with RA and the prevalence of clinically significant ILD was 9.8% in men and 6.8% in women in another study and studies by Natalini *et al* and Olson *et al*.[6 12] Despite this high heterogeneity in the prevalence of ILD in patients with RA (RAILD), it is now accepted that ILD is the most frequent lung manifestation in patients with RA.[13]

Patients with RAILD present higher morbidity and more affected quality of life than RA patients without ILD (RAnonILD).[6] However, the description of the characteristics in these patients is in general limited to small patient cohorts, and multicentric analyses with RAILD are still scarce. Hence, further research regarding the prevalence and characteristics of RA and RAILD is needed to better understand the disease and thus improve its management and prognosis. In recent years, the integration of natural language processing (NLP) and machine learning (ML) techniques have shown great potential in extracting valuable insights from electronic health records (EHRs) from patients with rheumatic diseases[14 15] and providing real-world evidence in the field. When applied to data, NLP algorithms can effectively extract and interpret unstructured clinical text, such as physician notes, pathology reports and treatment plans. By analysing this free-text information, NLP algorithms can identify and categorise specific populations, as well as clinical and treatment characteristics. The present real-world data (RWD) study aimed to provide the current scenario of patients with RA and RAILD. The specific objectives of this work were: (1) to estimate the nationwide prevalence of RA and RAILD in Spain and (2) to compare demographic, clinical characteristics and treatments between patients with RAILD and RAnonILD. In order to guarantee a real clinical setting, the information included in EHRs was analysed using EHRead,[16] a technology that applies NLP and ML, to analyse the free-text information.

METHODS

Study design, setting and participants

This was a multicentre, retrospective, observational case–control study based on clinical information captured in the EHRs of the participating hospitals. Data were collected from all available departments (including inpatient hospital, outpatient hospital and emergency room) from 1 January 2014 to 31 December 2019. The study was conducted in nine hospitals from the Spanish National Healthcare Network located in six different regions in Spain: Hospital Clínico Universitario (Comunidad Valenciana), Hospital Universitario y Politécnico la Fe (Comunidad Valenciana), Hospital Universitario de Fuenlabrada (Comunidad de Madrid), Hospital Universitario Infanta Leonor (Comunidad de Madrid), Hospital Universitario Puerta de Hierro (Comunidad de Madrid), Hospital Universitario de León (Castilla y León), Clínica Universitaria de Navarra (Navarra), Hospital Universitario Son Espases (Baleares), Hospital Universitario Vall d'Hebron (Cataluña). During the Hospital inventory and data integration process phase, it was noted that one of the participating hospitals did not fulfil the minimum data completeness criteria and was not included for further analysis.

The study population comprised all adult patients with available clinical information in any of the participating sites (ie, that attended the study hospitals) and had RA disease reported within the study period. Patients with RA with and without ILD were included in RAILD (ie, ILD within RA) and RAnonILD groups, respectively. Of note, we only included patients with at least one affirmed mention of RA or ILD, discarding those with only negated or speculated diagnosis in all the detections of the disease. The inclusion date was defined as the first detection of RA in the EHRs registered during routine clinical practice within the study period. Demographic, clinical and laboratory unstructured data (and structured, when available) were analysed at the time of inclusion within different time windows around the inclusion date. Regarding this, demographic characteristics and chest HRCT pattern of ILD were analysed from the first EHR available up to 2 months post-inclusion and comorbidities and blood tests were evaluated within a time window spanning from −6 to +2 months relative to the inclusion

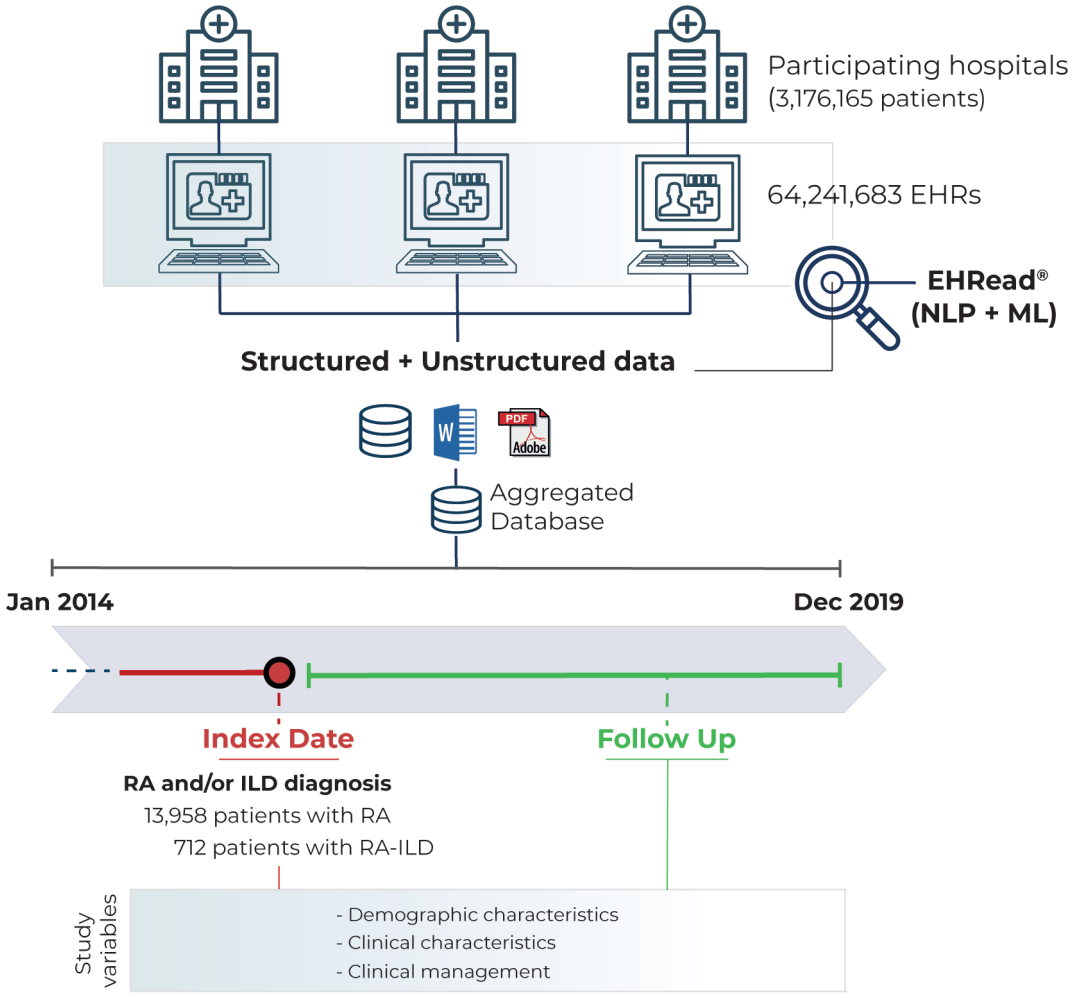

Figure 1 Study design. EHRead technology is a system based on natural language processing (NLP) that applies machine learning (ML) and deep learning to extract, analyse and interpret the free-text information written in millions of de-identified electronic health records (EHRs). The unstructured (and structured, if available), free-text information from EHRs from multiple participating sites was organised in study databases. Specific inclusion and exclusion criteria are specified to define the target population. The variables extracted from the database at different time points were organised and analysed to address multiple clinical questions. ILD, interstitial lung disease; RA, rheumatoid arthritis.

date. These distinct windows were used to maximise data availability across variables. For laboratory values, if multiple values were detected in the time window, only the closest to inclusion was retained. Treatment received (disease-modifying antirheumatic drugs (DMARDs), antifibrotics, immunosuppressors and glucocorticoids), healthcare resource utilisation and hospital mortality were also considered and analysed during follow-up in RAILD and RAnonILD. The follow-up period spanned from the start date of follow-up to the last available EHR for each patient within the study period (figure 1). The default follow-up start date was set as the date of inclusion. In patients who developed ILD post-inclusion, the follow-up period spanned from the ILD diagnosis to the last available EHR within the study period. To ensure fair comparisons between groups, the duration from inclusion to ILD diagnosis (or follow-up start date) in the RAILD group was quantified and the follow-up start date for RAnonILD patients was adjusted to randomly match a time point since inclusion from the RAILD group.

Extraction of information from EHRs

Date of birth and sex were extracted from structured data and age was computed based on birthdate (inclusion date – birthdate). Specific study variables, comorbidities (a detailed list of the assessed comorbidities is presented in online supplemental table 3), treatments and mortality were extracted from free text using the EHRead technology. It uses NLP and ML techniques for extracting free text from de-identified and processed EHRs. This technology has been used to extract unstructured clinical data and translate it into study databases.[17–19] The terminology used is based on SNOMED-CT and contains codes, concepts, synonyms, and definitions used in clinical documentation.[20 21]

Study variables—including RA, ILD and RAILD—were defined by the EHRead detection of keywords that referred to the presence of each characteristic. To generate SNOMED code lists for diagnoses of RA, and ILD subtypes, the SNOMED browser was searched for relevant concepts pertaining to medical

terms. The curated methodology approach used not only detects terms in the free text, but also their affirmation status. SNOMED code lists were reviewed by a consultant rheumatologist. Some variables were also derived from other variables. Leucocytosis was defined as leucocyte levels >12 000/mm^3. Rheumatoid factor (RF), and anti-citrullinated peptide antibodies (ACPA) positivity or negativity were defined based on specific cut-off values and as reported in free text. Any non-negated mention of death or synonyms was accounted as the moment of decease. EHRead technology and its internal validation have been described elsewhere.[16] A comprehensive list of these terms can be found in online supplemental table 1.

External validation of EHRead performance

Following the extraction of the free-text data from EHRs, it was determined whether the EHRead technology could accurately identify some of the main variables such as the target population and other important characteristics used in this investigation. This validation involved comparing the reading output of EHRead to a corpus of medical records that had been annotated by specialists acting as annotators in each participating institution (ie, the 'gold standard'). This validation methodology has been previously described.[16] Precision of variable detection for this study is shown in online supplemental table 2.

To reinforce the composition of the study population, ensuring it consists of patients with the specific disease under investigation, namely RA or ILD, we conducted an analysis to ascertain the disease, including presence of multiple mentions of RA or ILD in their EHRs, the utilisation of disease-specific treatments, the existence of disease activity indices within their EHRs and a positive result for RF or ACPA included as 'RA-support' subset (online supplemental tables 2 and 3).

Statistical analyses

Prevalence was estimated dividing the total number of patients with RA or RAILD by the total number of attended patients (online supplemental text 1). Both crude and age-adjusted prevalence were calculated. The frequency and prevalence of patients with ILD among patients with RA was also determined. Then, frequency was analysed as the proportion of patients with RA who also had an ILD diagnosis and prevalence as the average prevalence of RAILD throughout the study period, reflecting the dynamic patient population over time. Additionally, we conducted sensitivity analysis using the RA-support patients as the denominators for the descriptive analysis, and analysis changing the denominator to patients treated with DMARDs or glucocorticoids. Categorical variables were represented as frequencies, whereas numeric variables were summarised with median, IQR (Q1, Q3), mean, SD and the proportion of available data. The absence of information in EHRs was regarded as 'true zero' for binary variables and no further imputation was done. Fisher's exact tests, independent-samples T tests or Mann-Whitney U tests (ie, Wilcoxon tests) were used to compare groups based on categorical, normal numeric or non-normal numeric variables (respectively). Welch's adjustment was incorporated to t-tests for unequal variances. Normality was assessed with the Shapiro-Wilk test. Survival analyses were addressed through the Kaplan-Meier approach, and Cox proportional hazards regression models. The event measured was in-hospital mortality, from the time of inclusion (or ILD diagnosis, for patients with later ILD onset). Patients with no reported death were censored at the time of their last EHR available. Significance was defined as $p<0.05$ in two-tailed tests, with Benjamini and Hochberg adjustment for multiple hypothesis testing. Data was analysed and represented using 'R' software, V.4.0.2.

RESULTS

Population and prevalence of patients with RA and RAILD

The source population in the hospitals of interest comprised 3 176 165 patients, involving a total of 64 241 683 EHRs. The screening set-total number of patients who attended the hospital sites during the study period (2014–2019) included a total of 13 958 patients with a diagnosis of RA, and 712 patients were diagnosed with both RA and ILD (RAILD 5.1%) (figure 1). A detailed analysis on the number of EHRs per patient in our study population and by group can be seen in online supplemental table 3. Of note, 89.8% of the patients in the study population had at least one characteristic that supported an RA diagnosis, besides the mention of RA pathology in the EHRs.

The overall crude prevalence of RA in the geographical areas covered by the participating hospitals was estimated at 0.6% (600 cases per 100 000 individuals), and it varied considerably between the sexes (0.33% in men vs 0.83% in women). The estimated RAILD prevalence was 0.03% (30 cases per 100 000 individuals), with also a slight predominance in women (0.03% vs 0.02% in men) (online supplemental table 4). After weighing the prevalence according to the distribution of the population in Spain, we observed an overall age-adjusted prevalence of 0.53% and 0.02% for RA and RAILD, respectively (figure 2, online supplemental table 5). Of note, the average prevalence (95% CI) of ILD within the RA population (RAILD) in the study period was 4.5% (3.2% to 5.9%).

Characteristics of RAILD and RAnonILD patients at the time of inclusion

The demographic characteristics of patients with RAILD and RAnonILD and available information at the time of inclusion (n=13 246, 94.9%) are shown in table 1. RAILD were older with a higher percentage of smokers than RAnonILD (65 (±13) and 60 (±17) years, $p<0.001$; 65.6%

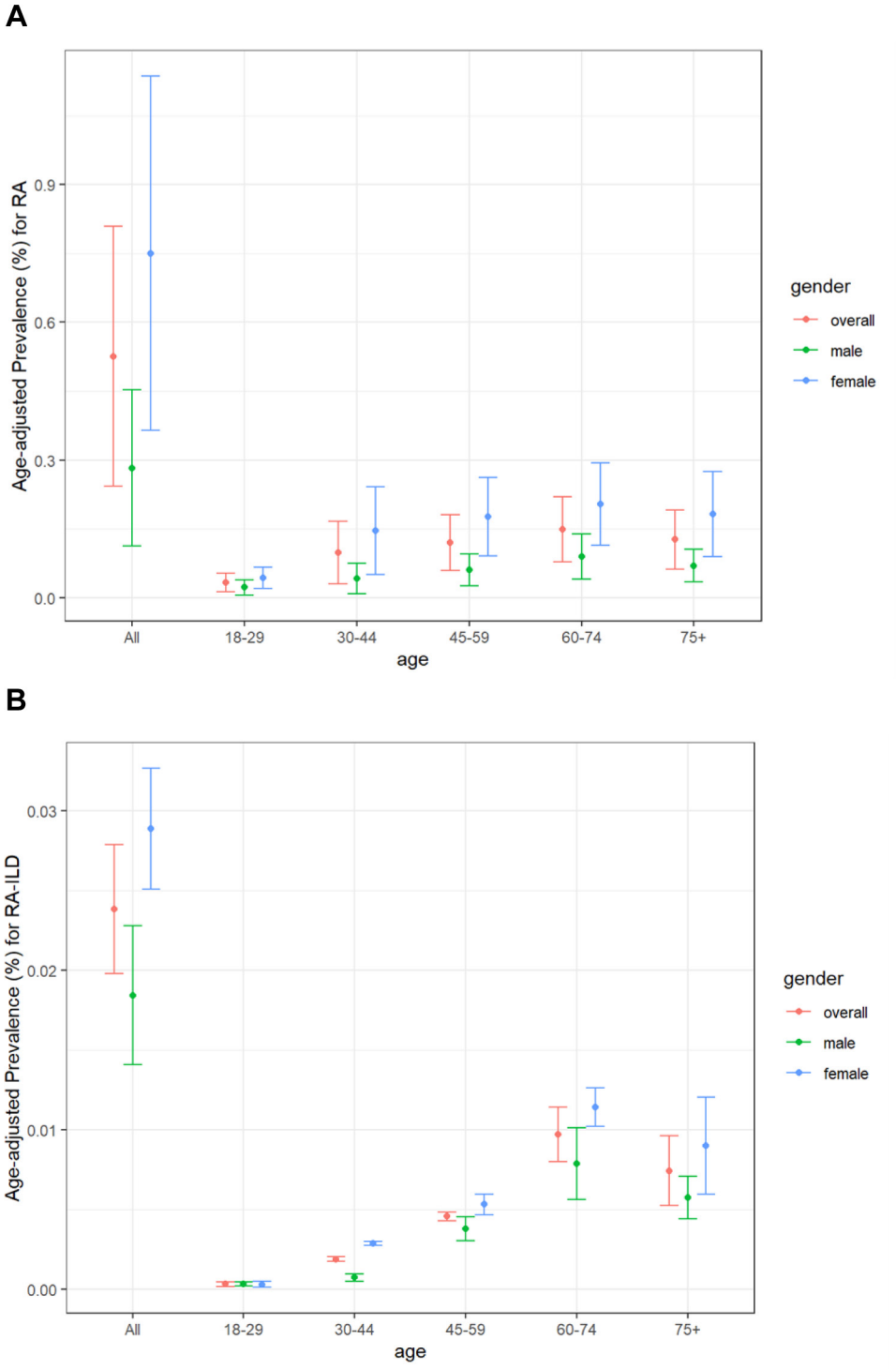

Figure 2 Age-adjusted prevalence for RA (A) and RAILD (B). ILD, interstitial lung disease; RA, rheumatoid arthritis.

vs 57.7%, p<0.001) with a predominance of female sex in both groups (61.7% vs 74.4%, p<0.001). Mean age at RA first mentioned for total RA population, RAILD and RAnonILD patients was 58 (±17), 61 (±15) and 58 (±18), respectively. Among patients with RAILD, age at ILD diagnosis was 64 (±14). The ILD subtype was specified in only 55.6% within the EHRs and the most reported were usual interstitial pneumonia (29.3%) and non-specific interstitial pneumonia (22.7%) (online supplemental table 6). ILD was detected in a median of 2 (0, 7) years after RA diagnosis (online supplemental table 6).

The most frequent comorbidities in both groups, RAILD and RAnonILD were hypertension (48.0% vs 36.6%, p<0.001) and dyslipidaemia (56.7% vs 46.6%, p<0.001). Most of the assessed comorbidities were more frequent in patients with RAILD as compared with RAnonILD (figure 3, online supplemental table 7). Of note, major comorbidities such as chronic obstructive

Table 1 Demographic characteristics at the time of inclusion

	RA (13 958)	RAILD (712)	RAnonILD (13 246)	P value	OR (95% CI)
Sex, n (%)					
Female	10290 (73.7)	439 (61.7)	9851 (74.4)	<0.001*	0.55 (0.47 to 0.65)
Male	3668 (26.3)	273 (38.3)	3395 (25.6)	<0.001*	1.80 (1.54 to 2.11)
Age, years					
N (%)	13958 (100)	712 (100)	13246 (100)		
Mean (SD)	60 (17)	65 (13)†	60 (17)	<0.001*	4.80 (3.77 to 5.83)‡
Median (Q1, Q3)	61 (49, 73)	65 (57, 75)	61 (48, 73)		
Tobacco, n (%)§	4581 (32.8)	343 (48.2)	4238 (32.0)	<0.001*	1.98 (1.69 to 2.31)
Smoker/ex-smoker	2670 (58.3)	225 (65.6)	2445 (57.7)	0.004*	1.40 (1.10 to 1.78)
Never smoker	1911 (41.7)	118 (34.4)	1793 (42.3)	0.004*	0.72 (0.56 to 0.91)
Alcohol, n (%)§	1029 (7.4)	50 (7.0)	979 (7.4)	0.769	0.95 (0.69 to 1.27)
Consumer/ex-consumer	316 (30.7)	19 (38.0)	297 (30.3)	0.272	1.41 (0.74 to 2.62)
Never consumer	713 (69.3)	31 (62.0)	682 (69.7)	0.272	0.71 (0.38 to 1.35)

Data extracted and analysed from the first EHR available to 2 months after the inclusion date. If multiple values were detected in the time window, only the closest to inclusion was retained.
*Statistical differences were considered significant when p<0.05 in two-tailed tests.
†Median (Q1, Q3) is preferred over mean (SD) for interpretation as the feature is non-normal.
‡Welch's t-test (difference of group means) were performed for statistical comparison of RAILD versus RAnonILD patients.
§Percentages calculated considering patients with available data.
EHR, electronic health record; ILD, interstitial lung disease; OR, OR (Fisher's test); RA, rheumatoid arthritis.

pulmonary disease (COPD) presented important differences between both groups RAILD and RAnonILD (OR, OR: 3.03; 95% CI 2.38 to 3.84)), as well as general infections (OR: 2.55; 95% CI 2.16 to 3.00), cardiovascular disease (OR: 2.16; 95% CI 1.80 to 2.58) and malignancies (OR: 2.02; 95% CI 1.62 to 2.50).

Table 2 shows the levels of inflammatory parameters related to RA at the time of inclusion. Higher concentrations of acute phase reactants as C-reactive protein (CRP) and erythrocyte sedimentation rate (ESR) were observed in patients with RAILD when compared with RAnonILD (median (Q1, Q3): 11.6 (2.8, 37) vs 5.5 (1.2, 21), p<0.004 and 36.5 (28.9) vs 27.8 (26.4), p=0.001, respectively). Other blood test results at the time of inclusion are detailed in online supplemental table 8. Briefly, 28.8% of patients with RAILD showed anaemia and 17.8% leucocytosis, while they were observed in 19.4% and 9.0% of the patients in the RAnonILD group. The autoantibodies status was unknown in over 75% of the patients. In those where it was reported we observed a higher frequency of RF (OR=0.55 (95% CI 0.46 to 0.65), p<0.001), and ACPA (OR=0.85 (95% CI 0.71 to 1.01), p=0.004) than in the RAnonILD group.

A sensitivity analysis using patients with at least one characteristic that supported an RA diagnosis found comparable results on the distribution of the main variables (online supplemental table 9).

Follow-up analysis of patients with RAILD and RAnonILD

Follow-up data in RAILD patients was longer than in RAnonILD (25±19 vs 19±18 months, p= <0.001). During the first 2 years of follow-up, RAILD patients had more treatment prescriptions as compared with RAnonILD (table 3). Thus, 44.7% of RAILD patients were prescribed biological DMARDs, while only 38.2% of the RAnonILD received these (p<0.001). The most used biological drugs in patients with RAILD were rituximab (9.8%), abatacept (8.7%) and adalimumab (6.9%). Other treatments as immunosuppressors (cyclophosphamide, mycophenolate and azathioprine), or glucocorticoids were also mainly used in patients with RAILD compared with RAnonILD (OR: 9.30 (95% CI 5.67 to 14.96); OR: 26.58 (95% CI 17.88 to 39.71); OR: 7.18 (95% CI 5.02 to 10.12) and OR: 4.83 (95% CI 4.13 to 5.66), all p<0.001). No significant differences were found in the use of methotrexate. Supplementary sensitivity analysis including only patients treated with DMARDs or corticoids are shown in online supplemental table 10.

Patients with RAILD had more hospitalisation during follow-up as compared with patients with RA (62.1% vs 30.4%, p<0.001), outpatient visits (86.7% vs 76.4%, p<0.001) and emergency visits (61.1% vs 42.8%, p<0.001 (online supplemental table 11). Finally, a higher percentage of in-hospital death was observed in patients with RAILD compared with RAnonILD, (15.4% vs 5.7%, p<0.001). Likewise, survival analyses demonstrated increased mortality in patients with RAILD compared with those with RAnonILD (figure 4). Specifically, the risk of death was considerably higher in the RAILD group, with an adjusted HR of 1.70; 95% CI 1.41 to 2.06, p<0.001.

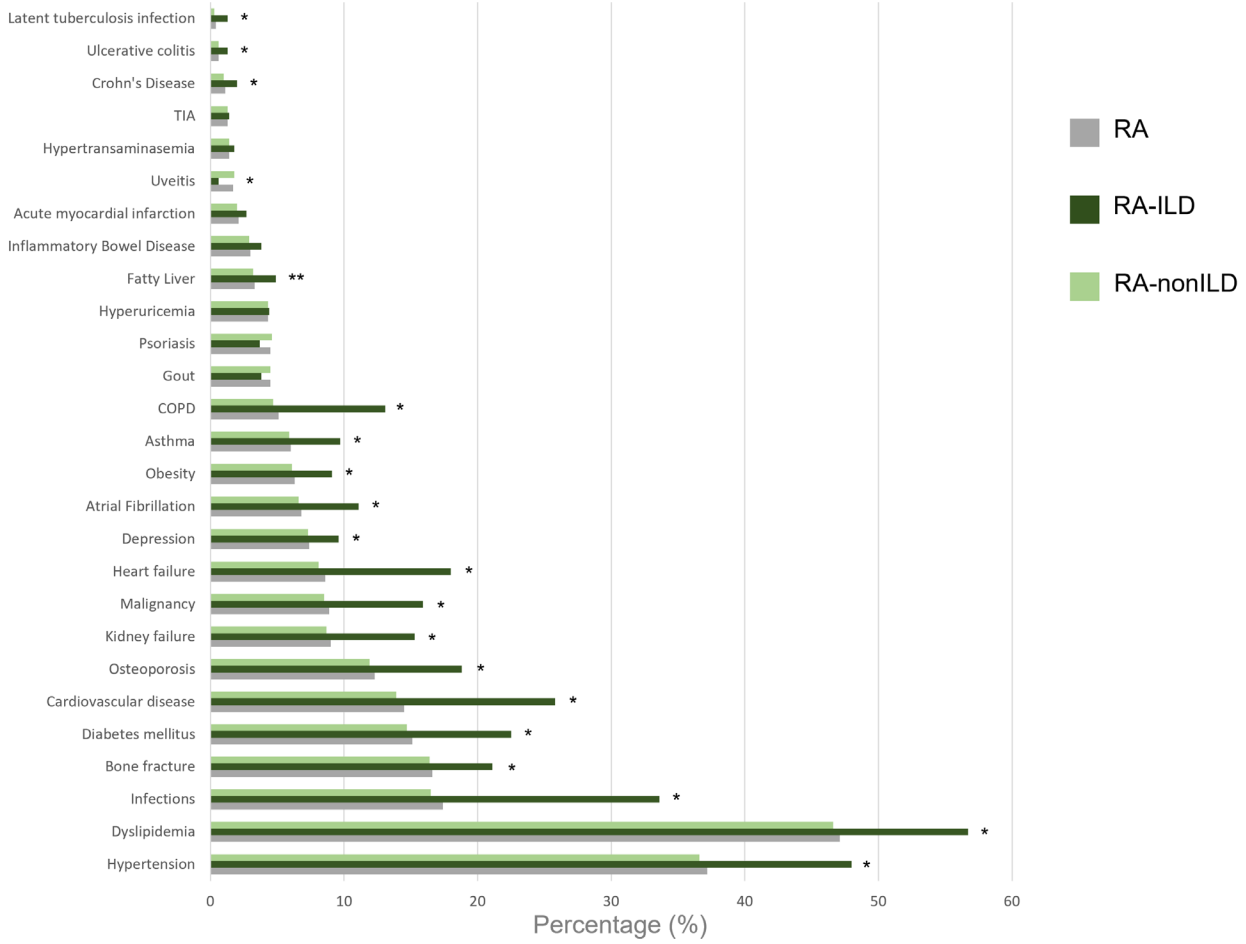

Figure 3 Frequency of comorbidities at the time of inclusion. COPD, chronic obstructive pulmonary disease; ILD, interstitial lung disease; RA, rheumatoid arthritis; TIA, transient ischaemic attack. *Statistical differences between RAILD and RAnonILD. Differences were considered significant when p<0.05 in two-tailed tests.

DISCUSSION

This study represents the first attempt to apply NLP and ML to extract and analyse real-life clinical information from patients with RA and RAILD. We identified 13 958 patients with RA diagnosis, among which 712 had RAILD. Our main results were: (1) Overall age-adjusted prevalence were 0.53% and 0.02% for RA and RAILD, respectively, and (2) Patients with RAILD when compared with RAnonILD were older, predominantly smokers and with more comorbidities and higher inflammatory parameters levels at the time of inclusion. Moreover, they used biological DMARDs more frequently and presented higher in-hospital mortality.

A recent meta-analysis based on a systematic review estimated a global prevalence of RA between 1980 and 2019 of 460 per 100 000 population.[2] Interestingly, when reviewing population-based studies, the mean point prevalence and the mean period prevalence were 0.56% (SD 0.51) and 0.51% (SD 0.35), respectively.[22] As per two national surveys based on random stratified multistage cluster sampling, Spain's RA prevalence ranges between 0.5% and 0.8%.[23 24] Despite the methodological differences of our study, we found comparable results to these reports, suggesting a firm backing to this novel approach.

According to our data, 1 out of 20 patients with RA presents ILD. Determining an accurate RAILD prevalence represents a great challenge mostly due to its relatively low frequency, as well as the lack of universal consensus in the whom (eg, asymptomatic vs symptomatic), the how (eg, pulmonary function test vs HR-CT) or when to assess for ILD. All the above mentioned is reflected in the wide heterogeneous prevalence reported in current literature, which ranges from 1% to 61%.[25–30] Large studies based on administrative structured data sets built on the International Classification of Disease coding, estimates RAILD prevalence between 2% and 6% among the RA population.[4 26 31] Yet the complexity claims-based algorithm limits the ability to accurately identify RAILD (positive predictive value 44–72%).[32] By using NLP, we were able to analyse unstructured and structured (when available), with a precision of 79.4% and 76.4% for ILD and RA, respectively. Moreover, 90% of the patients with RA in our study had at least one characteristic that supported an RA diagnosis, on top of the mention of RA in the EHRs. Besides, we have performed sensitivity analysis with comparable results to the original findings, which reinforces the robustness and reliability of our research approach. Twenty years ago, the EMECAR (Estudio de

Table 2 Autoantibodies status and inflammatory serum parameters at the time of inclusion

	RA (13958)	RAILD (712)	RAnonILD (13246)	P value	OR (95% CI)
Leucocytes, n (%)	5219 (37.4)	310 (43.5)	4909 (37.1)		
Mean (SD), 1000/mm^3	12.2 (42.9)	15.9 (60.7)*	12 (41.5)*	0.265	3.90 (–2.98 to 10.78)†
Median (Q1, Q3), 1000/mm^3	8 (6.1, 10.8)	8.7 (6.4, 11.5)	8 (6.1, 10.7)		
Leucocytosis,‡ n (%)	1320 (9.5)	127 (17.8)	1193 (9)	<0.001§	2.19 (1.78 to 2.69)
CRP, n (%)	6194 (44.4)	350 (49.2)	5844 (44.1)		
Mean (SD), mg/L	30.3 (71.2)	40.8 (71.1)*	29.6 (71.1)	0.004§	11.13 (3.45 to 18.80)†
Median (Q1, Q3), mg/L	5.8 (1.3, 22)	11.6 (2.8, 37)	5.5 (1.2, 21)		
CRP >5 mg/L, n (%)	3620 (25.9)	256 (36)	3364 (25.4)	<0.001§	1.65 (1.40 to 1.94)
ESR, n (%)	5222 (37.4)	291 (40.9)	4931 (37.2)		
Mean (SD), mm/hour	28.2 (26.6)	36.5 (28.9)*	27.8 (26.4)*	<0.001§	8.77 (5.35 to 12.18)¶
Median (Q1, Q3), mm/hour	20 (8, 40)	32 (15, 50)	20 (8, 39.6)		
ESR >20 mm/hour, n (%)	2984 (21.4)	210 (29.5)	2774 (20.9)	<0.001§	1.58 (1.33 to 1.87)
RF, n (%)	2860 (20.5)	219 (30.8)	2641 (19.9)		
Mean (SD), U/mL	123.3 (239)*	165.7 (308.1)*	119.7 (232)*	0.321	1.00 (–1.00 to 6.50)†
Median (Q1, Q3), U/mL	36 (10, 111)	43 (10, 172)	36 (10, 107)		
Positive RF, n (%)**	2022 (14.5)	162 (22.8)	1860 (14)	<0.001§	1.80 (1.49 to 2.17)
Negative RF, n (%)**	1100 (7.9)	78 (11)	1022 (7.8)	0.003§	1.47 (1.14 to 1.88)
Unknown RF, n (%)**	10 836 (77.6)	472 (66.7)	10 364 (78.2)	<0.001§	0.55 (0.46 to 0.65)
ACPA, n (%)	3362 (24.1)	193 (27.1)	3169 (23.9)		
Mean (SD), U/mL	222 (443.2)*	274.1 (515.6)*	218.8 (438.3)*	0.143	1.60 (–0.25 to 9.60)†
Median (Q1, Q3), U/mL	47.4 (3, 250)	73 (4.6, 290.3)	45.7 (3, 250)		
Positive ACPA, n (%)††	2083 (14.9)	134 (18.8)	1949 (14.7)	0.004§	1.34 (1.10 to 1.63)
Negative ACPA, n (%)††	1338 (9.6)	62 (8.7)	1276 (9.6)	0.472	0.89 (0.67 to 1.17))
Unknown ACPA, n (%)††	10 537 (75.5)	516 (72.5)	10 021 (75.7)	0.060	0.85 (0.71 to 1.01)

Data extracted and analysed considering a window of (–6, +2 months) around the inclusion date. If multiple values were detected in the time window, only the closest to inclusion was retained.
*Median (Q1, Q3) is preferred over mean (SD) for interpretation as the feature is non-normal.
†Wilcoxon test (difference of location) was performed for statistical analysis of Leucocytes, CRP, RF and ACPA for RAILD versus RAnonILD patients.
‡Leucocytosis was defined as leucocytes levels >12 000/mm^3.
§Statistical differences were considered significant when p<0.05 in two-tailed tests.
¶Welch t-test was performed for statistical analysis of ESR.
**RF positive in free text or equal to or above 20; RF negative in free text or below 20.
††ACPA positive in free text or equal to or above 30; ACPA negative in free text or below 30.
ACPA, anticitrullinated protein antibody; CRP, C-reactive protein; ESR, erythrocyte sedimentation rate; ILD, interstitial lung disease; OR, OR (Fisher's test); RA, rheumatoid arthritis; RF, rheumatoid factor.

la Morbilidad y Expresión Clínica de la Artritis Reumatoide) study reported a rate of 3.7% of ILD (3.2%) among patients with RA in Spain. Differences in methodology might explain the discrepancies with our findings, but as some studies[4 12 31] have suggested, a raise in RAILD prevalence cannot be ruled out. A greater awareness on the diagnosis of RAILD by physicians, as well as the change in diagnostic criteria for RA may have influenced a greater identification of patients with RA, and therefore with RAILD.[33]

In general, patients with RA are at higher risk of developing certain comorbidities when compared with the general population.[34] Even so as in our findings, it has been described that RAILD are a more vulnerable subpopulation, with an increased morbidity compared with patients with RA without ILD.[4 35] Within our cohort, most of the assessed comorbidities were more frequent in the RAILD group, but it was more evident in COPD, infections, cardiovascular diseases and malignancies. The impact of the latter on the disease may be reflected in an increased mortality in patients with RAILD compared with those with RAnonILD with an almost twofold mortality risk in our survival analysis. Nearly 70% of excess deaths in RA are attributable to cardiopulmonary disease, and it has been suggested that ILD is the cause of death most strongly associated with RA.[36] We must remark that due

Table 3 Disease-related treatment during the follow-up period

	RA (13 958)	RAILD (712)	RAnonILD (13 246)	P value	OR (95% CI)
Conventional synthetic DMARDs, n (%)	5381 (38.6)	318 (44.7)	5063 (38.2)	<0.0.01*	1.30 (1.12 to 1.52)
Methotrexate	4994 (35.8)	271 (38.1)	4723 (35.7)	0.199	1.11 (0.95 to 1.30)
Leflunomide	1014 (7.3)	117 (16.4)	897 (6.8)	<0.001*	2.71 (2.18 to 3.35)
Sulfasalazine	117 (0.8)	13 (1.8)	104 (0.8)	0.009*	2.35 (1.20 to 4.22)
Biological DMARDs, n (%)	2257 (16.2)	194 (27.2)	2063 (15.6)	<0.001*	2.03 (1.70 to 2.42)
TNF inhibitors	1853 (13.3)	108 (15.2)	1745 (13.2)	0.126	1.18 (0.95 to 1.46)
Adalimumab	766 (5.5)	49 (6.9)	717 (5.4)	0.108	1.29 (0.94 to 1.75)
Certolizumab pegol	179 (1.3)	5 (0.7)	174 (1.3)	0.227	0.53 (0.17 to 1.27)
Etanercept	890 (6.4)	41 (5.8)	849 (6.4)	0.529	0.89 (0.63 to 1.23)
Golimumab	178 (1.3)	6 (0.8)	172 (1.3)	0.389	0.65 (0.23 to 1.44)
Infliximab	251 (1.8)	28 (3.9)	223 (1.7)	<0.001*	2.39 (1.54 to 3.58)
Other MoA	647 (4.6)	124 (17.4)	523 (3.9)	<0.001*	5.13 (4.11 to 6.36)
Abatacept	347 (2.5)	62 (8.7)	285 (2.2)	<0.001*	4.34 (3.20 to 5.80)
Rituximab	332 (2.4)	70 (9.8)	262 (2)	<0.001*	5.40 (4.04 to 7.15)
Tocilizumab	386 (2.8)	34 (4.8)	352 (2.7)	0.002*	1.84 (1.24 to 2.64)
Targeted synthetic DMARDs, JAK inhibitors, n (%)	76 (0.5)	2 (0.3)	74 (0.6)	0.439	0.50 (0.06 to 1.88)
Tofacitinib	56 (0.4)	2 (0.3)	54 (0.4)	1	0.69 (0.08 to 2.62)
Baricitinib	24 (0.2)	0 (0)	24 (0.2)	0.633	0.00 (0.00 to 3.09)
Filgotinib	1 (0)	0 (0)	1 (0)	1	0.00 (0.00 to 717.43)
Upadacitinib	1 (0)	0 (0)	1 (0)	1	0.00 (0.00 to 717.43)
Antifibrotics, n (%)	–	12 (1.7)	–	–	–
Nintedanib	–	5 (0.7)	–	–	–
Pirfenidone	–	8 (1.1)	–	–	–
Immunosuppressors, n (%)					
Cyclophosphamide	86 (0.6)	28 (3.9)	58 (0.4)	<0.001*	9.30 (5.67 to 14.96)
Mycophenolate	113 (0.8)	64 (9)	49 (0.4)	<0.001*	26.58 (17.88 to 39.71)
Azathioprine	184 (1.3)	49 (6.9)	135 (1)	<0.001*	7.18 (5.02 to 10.12)
Glucocorticoids, n (%)	2878 (20.6)	377 (52.9)	2501 (18.9)	<0.001*	4.83 (4.13 to 5.66)

Data extracted and analysed between the time of inclusion and 2 years of follow-up.
*Statistical differences were considered significant when p<0.05 in two-tailed tests.
DMARDs, disease-modifying antirheumatic drugs; ILD, interstitial lung disease; JAK, Janus kinase; MoA, mechanism of action; OR, OR (Fisher's tests); RA, rheumatoid arthritis; TNF, tumour necrosis factor.

to the nature of our data and the study design, determining death causality was not feasible.

Among the known risk factors of ILD in patients with RA, elderly, male sex, a history of smoking, extra-articular manifestations, auto-antibodies as well as RA disease activity have been described.[28 37–39] Moreover, some factors, such as relative diffusion capacity of the lung for carbon monoxide decline or RF positivity have been shown as predictors of mortality in patients with RAILD.[40] Interestingly, patients with RAILD in our study presented an older age, a greater male-to-female ratio, more smoking habit and a higher frequency of RF and ACPA antibodies as compared with RAnonILD. It is worth mentioning that the findings in antibodies were well below expectations. They are structured data not usually reflected in the EHR and therefore not detectable by NLP which could have explained suboptimal detection metrics. RAILD patients also had higher inflammatory load directly detected as an increase in acute phase reactants such as CRP and ESR and indirectly by the greater use of RA drugs, including DMARDs and glucocorticoids. These results suggest that the inflammatory activity of RA could be an independent factor in the development of RAILD, which is aligned with recent studies that show that active RA was associated with an increased risk of developing RAILD.[41] Our study was not designed to examine the effect of treatments on the development and evolution of ILD. Even so, we consider that adequate control of disease activity through treatments, including a good safety-efficacy balance, is critical in these patients.

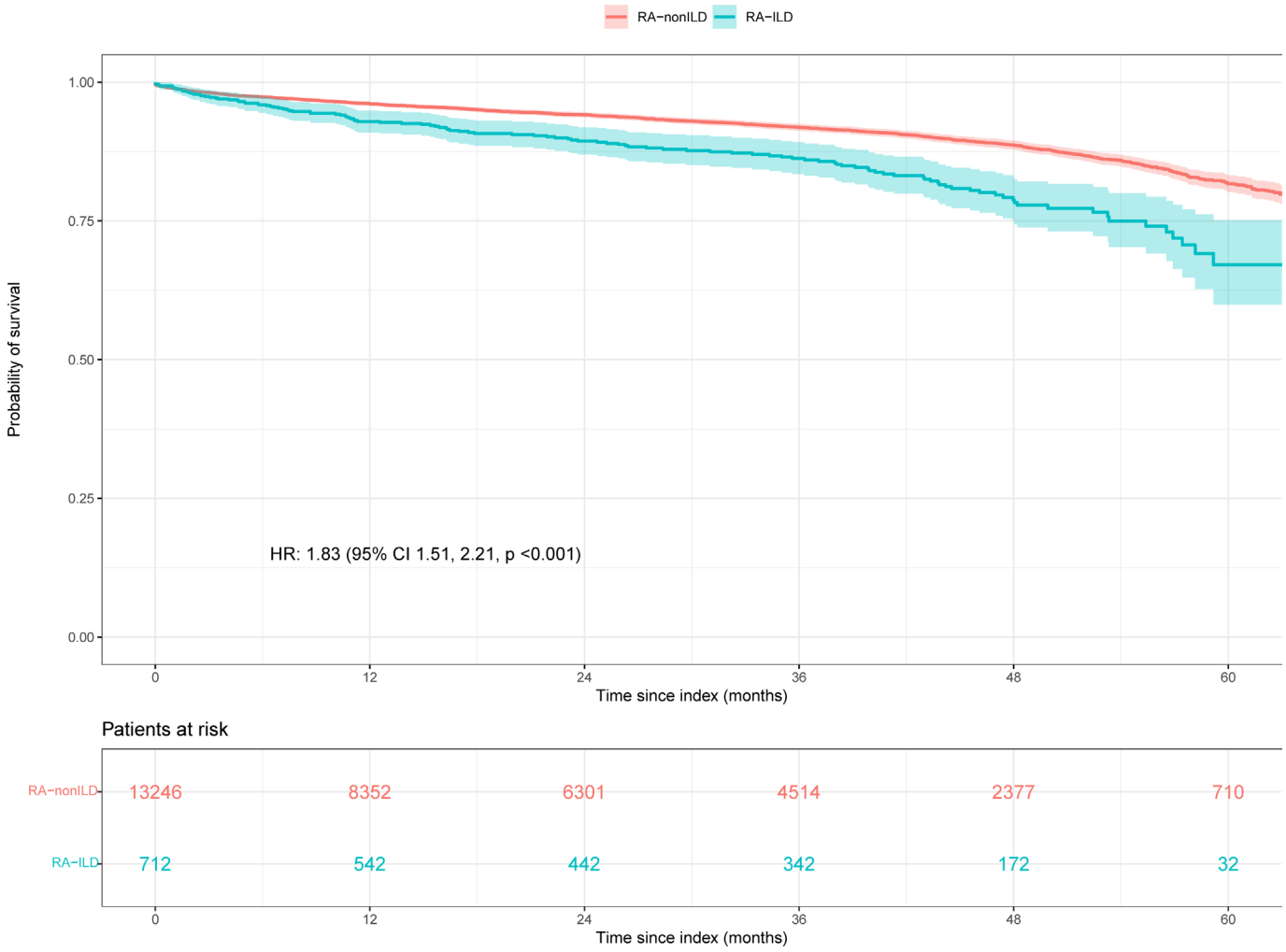

Figure 4 In-hospital mortality. Kaplan-Meier survival curves of RAILD and RAnonILD patients. Patients with no reported death were censored at the time of their last EHR available. EHRs, electronic health records; ILD, interstitial lung disease; RA, rheumatoid arthritis

Our study presents some limitations. First, results rely on the actual information reflected by physicians in clinical practice, as well as on the availability of EHRs. In this regard, full structured data from radiology reports, lung function tests, laboratory tests among others could not be assessed, and information concerning these fields may be limited. In addition, not all variables included in the study were available in all records, such as antibodies or disease activity. Regarding missing data for binary variables, we interpreted the absence of information as an absence of the characteristic given the inherent challenges of distinguishing between a non-present characteristic and an unreported or unevaluated one. Then, with our methodology we are able to promote a high precision (ie, positive predictive value), reducing the number of false detections to ensure that the information we captured was correct. Besides, the efficiency of the models used to read and interpret the data still have room for improvement, which may lead to misclassifications of certain variables, particularly in small groups; however, these are intrinsic characteristics of RWD studies, which are counterbalanced by a great number of data and patients studied. Thus, the lack of standardisation in terminology, omitted information or misuse of sections in the records represent a methodological limitation. Additionally, even if multiple records are available for a given patient, the availability of the desired variables in those records cannot be guaranteed. However, our multicentre approach involving diverse hospital sites and departments optimises an accurate and representative data retrieval letting us estimate for the very first time in a representative cohort of patients the Spanish RA and RAILD prevalence. Also, these results may not be generalisable to other populations. Additionally, although we investigated associations of factors RAILD at inclusion, these results should be interpreted with caution since the study was not designed for comparative safety for RAILD. As an example, the associations of medications and glucocorticoids with RAILD may be indicators of RA severity rather than directly affecting RAILD risk. As in all observational studies, residual confounding from unmeasured factors is possible. Additionally, the analysis of in-hospital death or prevalence may be subjected to survival bias since the study does include both RA

incident and prevalent patients. Nonetheless, we have tackled this limitation in the heterogeneous population of RA by adjusting the follow-up start date for RAnonILD patients to randomly match a time point since inclusion from the RAILD group. Finally, in our study patients with RAILD had both conditions, but the use of this nomenclature did not imply causality between RA and ILD.

CONCLUSION

In conclusion, by analysing readily available information in the EHRs using NLP and ML technologies we were able to characterise and to estimate the prevalence of RA and RAILD. RAILD subpopulation represents an older and more comorbid group with higher inflammatory burden, maybe related to an increased mortality compared with RAnonILD. Consequently, our results suggest that RAILD patients are a vulnerable population in whom closely monitoring during their follow-up should be recommended and an efficacy/safety balance when prescribing treatments is considered. An RWD-based understanding of the diseases could help clinicians to better adapt their treatment interventions in the future. Future studies should investigate whether RA-related medications may be associated with RAILD progression while appropriately accounting for confounding by indication or using randomised controlled trials.

Author affiliations
[1]Reumathology Department, Hospital Politécnico y Universitario La Fe, Valencia, Spain
[2]Rheumatology Department, Hospital Universitari Vall d'Hebron, Barcelona, Spain
[3]Rheumatology Department, Hospital Infanta Leonor, Madrid, Spain
[4]Pneumology Department, Hospital Infanta Leonor, Madrid, Spain
[5]Rheumatology Department, Hospital Universitario Son Espases, Palma, Spain
[6]Pneumology Department, Hospital Universitario Son Espases, Palma, Spain
[7]Rheumatology Department, Hospital Universitario Central de la Defensa Gómez Ulla, Madrid, Spain
[8]Rheumatology, Hospital Universitario Fuenlabrada, Madrid, Spain
[9]Rheumatology Department, Hospital Clinico Universitario, Valencia, Spain
[10]Rheumatology Department, Hospital Clínico Universitario, Valencia, Spain
[11]Pneumology Department, Hospital Clinico Universitario, Valencia, Spain
[12]Rheumatology Department, Clínica Universidad de Navarra, Pamplona, Spain
[13]Rheumatology Department, Clinica Universidad de Navarra, Pamplona, Spain
[14]Pneumology Department, Clinica Universidad de Navarra, Pamplona, Spain
[15]Rheumatology Department, Hospital Universitario Puerta de Hierro Majadahonda, Madrid, Spain
[16]Rheumatology Department, Complejo Asistencial Universitario de Leon, León, Spain
[17]Rheumatology Department, Complejo Asistencial Universitario de Leon, Leon, Spain
[18]Pneumology Department, Complejo Asistencial Universitario de Leon, Leon, Spain
[19]Medical Department, Savana Research SL, Madrid, Spain
[20]Health Economics and Outcomes Research, Bristol-Myers Squibb Company, Madrid, Spain
[21]Medical Department, Bristol-Myers Squibb Company, Madrid, Spain

Twitter Raul Castellanos-Moreira @raul_cast_morei

Acknowledgements The authors would like to express their gratitude to the support of Savana, including Victor Fanjul and Eva Castillo for their support with the analysis and interpretation of data; Eduard Sarró for his assistance with medical writing; Jose Aquino, Sebastian Menke and Ignacio Salcedo for their contributions to the natural language processing of this study; Judith Marín and Miren Taberna for their scientific support. The authors would also like to thank the BMS medical scientific liaisons Luz Ariza, Elena Sayagues and Nicolas de Miguel for their support along the study.

Contributors Contributorship: substantial contribution that qualify for authorship for the following: Conception or design: JARI, ET-A, MLL, LC, LL, BL-M, LS-F, RVC, PA, IDIMB, BS, EO, ED, JLA, DB, DV, SLV, RC-M. Data acquisition: JARI, LC, LL, JF-M, BN, LS-F, NCT, JR, AC, ALR, EB, DB. Data analysis: JARI, ET-A, MLL, BL-M, RVC, PA, DB, DV, SLV, RC-M. Data interpretation: JARI, ML- L, LC, LL, JF-M, B-N, LS-F, RVC, IDIMB, NCT, BS, EO, JR, AC, ED, ALR, EB, JLA, DB, DV, SLV, RC-M.RC-M accepts full responsibility for the finished work and/or the conduct of the study, had access to the data, and controlled the decision to publish.

Funding The RA-WILD study was fully supported by Bristol-Myers Squibb Company. Award/grant number not applicable.

Competing interests JARI has received support for research grants or contracts from BMS. ET-A has received payment honoraria for lectures, presentations, speakers bureaus, manuscript writing or educational events from BMS, and support for attending meetings from Boehringer Ingelheim, Nordic Pharma and Merck Sharp & Dohme. JF-M has received payment honoraria for lectures, presentations, speakers bureaus, manuscript writing or educational events from BMS, Amgen, Lilly, Galápagos, Boehringer Ingelheim, Novartis and also support for attending meetings from Boehringer Ingelheim. LS-F has received payment honoraria for lectures, presentations, speakers bureaus, manuscript writing or educational events from Novartis, BMS, Lilly, support for attending meetings from Pfizer, Lilly and Novartis and also has participated on Data Safety Monitoring Board or Advisory Board for Novartis, Merck Sharp & Dohme and Sanofi. BS has received consulting fees from Boehringer Ingelheim, payment honoraria for lectures, presentations, speakers bureaus, manuscript writing or educational events from Roche and Boehriner Ingelheim, support for attending meetings from Boehringer Ingelheim, Roche and AstraZeneca. JLA has received consulting fees from AbbVie, Amgen, AstraZeneca, Biogen, Cellgene, Celltion, Fresenius-Kabi, Galápagos, Gebro, GSK, Merck Sharp & Dohme, Pfizer, Regeneron, UCB, payment honoraria for lectures, presentations, speakers bureaus, manuscript writing or educational events from AbbVie, Antares, Biogen, GSK, Janssen, Merck Sharp & Dohme, Lilly, Nordic, Novartis, Sanofi, UCB and also support for attendings from AbbVie, Gebro, Janssen, Merck Sharp & Dohme, Nordic, Novartis and Pfizer. EB has received payment honoraria for educational events from Boehringer Ingelheim, support for attending meetings and/or travel from Boehringer Ingelheim and Janssen. DB currently works at Savana Research, has received grants or contracts from Novartis, payment honoraria for lectures, presentations, speakers bureaus, manuscript writing or educational events from Janssen and AbbVie, support for attending meetings from UCB, Novartis and AbbVie. DV works at BMS Company and own stocks as BMS employee. SLV works at BMS Company and own stocks as BMS employee. RACM works at BMS Company. The present manuscript was fully supported by Bristol-Myers Squibb.

Patient consent for publication Not applicable.

Ethics approval The protocol, amendments, and subject informed consent received appropriate approval by the Investigation Ethics Committees of the Baleares Islands (CIEB), or other applicable review board as required by local law prior to initiation of study at the site (ethics approval ID: CIEB-IB IB4369-20 EPA). All procedures and analysis adhered to legal and regulatory standards for good research practices as outlined in the most recent edition of the Helsinki Declaration. Patient agreement was not needed for this study because all information was gathered from anonymous electronic health recordsEHRs. To ensure confidentiality, EHRs were pseudonymizedpseudonymised and subsequently aggregated, before being translated into an anonymizedanonymised Study Database. Patients and the public were not formally consulted regarding this study.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement All data relevant to the study are included in the article or uploaded as supplementary information.

ORCID iD
Raul Castellanos-Moreira http://orcid.org/0000-0002-4104-4101

REFERENCES

1 Smolen JS, Aletaha D, Barton A, *et al*. Rheumatoid arthritis. *Nat Rev Dis Primers* 2018;4:18001.
2 Almutairi K, Nossent J, Preen D, *et al*. The global prevalence of rheumatoid arthritis: a meta-analysis based on a systematic review. *Rheumatol Int* 2021;41:863–77.
3 Bongartz T, Nannini C, Medina-Velasquez YF, *et al*. Incidence and mortality of interstitial lung disease in rheumatoid arthritis: a population-based study. *Arthritis Rheum* 2010;62:1583–91.
4 Hyldgaard C, Hilberg O, Pedersen AB, *et al*. A population-based cohort study of rheumatoid arthritis-associated interstitial lung disease: comorbidity and mortality. *Ann Rheum Dis* 2017;76:1700–6.
5 Jacob J, Hirani N, van Moorsel CHM, *et al*. Predicting outcomes in rheumatoid arthritis related interstitial lung disease. *Eur Respir J* 2019;53:1800869.
6 Natalini JG, Swigris JJ, Morisset J, *et al*. Understanding the determinants of health-related quality of life in rheunatoid arthritis-associated Intestitial lung disease. *Respir Med* 2017;127:1–6.
7 Valerio F, Daccord C, Letovanec I, *et al*. Rheumatoid arthritis-associated interstitial lung disease: new genetic data and therapeutic perspectives. *Rev Med Suisse* 2019;15:536–41.
8 Brito Y, Glassberg MK, Ascherman DP. Rheumatoid arthritis-associated interstitial lung disease: current concepts. *Curr Rheumatol Rep* 2017;19:12.
9 Suzuki A, Ohosone Y, Obana M, *et al*. Cause of death in 81 autopsied patients with rheumatoid arthritis. *J Rheumatol* 1994;21:33–6.
10 Stack BH, Grant IW. Rheumatoid interstitial lung disease. *Br J Dis Chest* 1965;59:202–11.
11 Frank ST, Weg JG, Harkleroad LE, *et al*. Pulmonary dysfunction in rheumatoid disease. *Chest* 1973;63:27–34.
12 Olson AL, Swigris JJ, Sprunger DB, *et al*. Rheumatoid arthritis-interstitial lung disease-associated mortality. *Am J Respir Crit Care Med* 2011;183:372–8.
13 Kadura S, Raghu G. Rheumatoid arthritis-interstitial lung disease: manifestations and current concepts in pathogenesis and management. *Eur Respir Rev* 2021;30:160.
14 Maarseveen TD, Meinderink T, Reinders MJT, *et al*. Machine learning electronic health record identification of patients with rheumatoid arthritis: algorithm pipeline development and validation study. *JMIR Med Inform* 2020;8:e23930.
15 Humbert-Droz M, Izadi Z, Schmajuk G, *et al*. Development of a natural language processing system for extracting rheumatoid arthritis outcomes from clinical notes using the national rheumatology informatics system for effectiveness registry. *Arthritis Care Res (Hoboken)* 2023;75:608–15.
16 Canales L, Menke S, Marchesseau S, *et al*. Assessing the performance of clinical natural language processing systems: development of an evaluation methodology. *JMIR Med Inform* 2021;9:e20492.
17 Ancochea J, Izquierdo JL, Soriano JB. Evidence of gender differences in the diagnosis and management of coronavirus disease 2019 patients: an analysis of electronic health records using natural language processing and machine learning. *J Womens Health (Larchmt)* 2021;30:393–404.
18 Izquierdo JL, Ancochea J, Soriano JB, *et al*. Clinical characteristics and prognostic factors for intensive care unit admission of patients with COVID-19: retrospective study using machine learning and natural language processing. *J Med Internet Res* 2020;22:e21801.
19 Graziani D, Soriano JB, Del Rio-Bermudez C, *et al*. Characteristics and prognosis of COVID-19 in patients with COPD. *J Clin Med* 2020;9:10.
20 Espinosa L, Tello J, Pardo A, *et al*. *Savana: a global information extraction and terminology expansion framework in the medical domain,* vol 57. 2016: 8.
21 Benson T. *Principles of health. Interoperability HI7And SNOMED*. New York: Springer London Dordrecht Heidelberg, 2010: 1–271.
22 Almutairi KB, Nossent JC, Preen DB, *et al*. The prevalence of rheumatoid arthritis: a systematic review of population-based studies. *J Rheumatol* 2021;48:669–76.
23 Silva-Fernández L, Macía-Villa C, Seoane-Mato D, *et al*. The prevalence of rheumatoid arthritis in Spain. *Sci Rep* 2020;10:21551.
24 Carmona L, Villaverde V, Hernández-García C, *et al*. The prevalence of rheumatoid arthritis in the general population of Spain. *Rheumatology (Oxford)* 2002;41:88–95.
25 Atienza-Mateo B, Remuzgo-Martínez S, Mora Cuesta VM, *et al*. The spectrum of interstitial lung disease associated with autoimmune diseases: data of a 3.6-year prospective study from a referral center of interstitial lung disease and lung transplantation. *JCM* 2020;9:1606.
26 Sparks JA, Jin Y, Cho S-K, *et al*. Prevalence, incidence and cause-specific mortality of rheumatoid arthritis-associated interstitial lung disease among older rheumatoid arthritis patients. *Rheumatology (Oxford)* 2021;60:3689–98.
27 Esposito AJ, Sparks JA, Gill RR, *et al*. Screening for preclinical parenchymal lung disease in rheumatoid arthritis. *Rheumatology (Oxford)* 2022;61:3234–45.
28 Juge P-A, Granger B, Debray M-P, *et al*. A risk score to detect Subclinical rheumatoid arthritis-associated interstitial lung disease. *Arthritis Rheumatol* 2022;74:1755–65.
29 Garrote-Corral S, Silva-Fernández L, Seoane-Mato D, *et al*. Screening of interstitial lung disease in patients with rheumatoid arthritis: a systematic review. *Reumatol Clin (Engl Ed)* 2022;18:587–96.
30 Bonilla Hernán MG, Gómez-Carrera L, Fernández-Velilla Peña M, *et al*. Prevalence and clinical characteristics of symptomatic diffuse interstitial lung disease in rheumatoid arthritis in a Spanish population. *Rev Clin Esp (Barc)* 2022;222:281–7.
31 Raimundo K, Solomon JJ, Olson AL, *et al*. Rheumatoid arthritis-interstitial lung disease in the United States: prevalence incidence, and healthcare costs and mortality. *J Rheumatol* 2019;46:360–9.
32 Cho S-K, Doyle TJ, Lee H, *et al*. Validation of claims-based algorithms to identify interstitial lung disease in patients with rheumatoid arthritis. *Semin Arthritis Rheum* 2020;50:592–7.
33 Neogi T, Aletaha D, Silman AJ, *et al*. The 2010 American college of rheumatology/European league against rheumatism classification criteria for rheumatoid arthritis: phase 2 methodological report. *Arthritis Rheum* 2010;62:2582–91.
34 Kronzer VL, Crowson CS, Sparks JA, *et al*. Comorbidities as risk factors for rheumatoid arthritis and their accrual after diagnosis. *Mayo Clin Proc* 2019;94:2488–98.
35 O'Dwyer DN, Armstrong ME, Cooke G, *et al*. Rheumatoid arthritis (RA) associated interstitial lung disease (ILD). *Eur J Intern Med* 2013;24:597–603.
36 Johnson TM, Yang Y, Roul P, *et al*. A narrowing mortality gap: temporal trends of cause-specific mortality in a national matched cohort study in US veterans with rheumatoid arthritis. *Arthritis Care Res (Hoboken)* 2023;75:1648–58.
37 Castellanos-Moreira R, Rodríguez-García SC, Gomara MJ, *et al*. Anti-carbamylated proteins antibody repertoire in rheumatoid arthritis: evidence of a new autoantibody linked to interstitial lung disease. *Ann Rheum Dis* 2020;79:587–94.
38 Samhouri BF, Vassallo R, Achenbach SJ, *et al*. Incidence, risk factors, and mortality of clinical and subclinical rheumatoid arthritis-associated interstitial lung disease: a population-based cohort. *Arthritis Care Res (Hoboken)* 2022;74:2042–9.
39 Albrecht K, Strangfeld A, Marschall U, *et al*. Interstitial lung disease in rheumatoid arthritis: incidence, prevalence and related drug prescriptions between 2007 and 2020. *RMD Open* 2023;9:e002777.
40 Hyldgaard C, Ellingsen T, Hilberg O, *et al*. Rheumatoid arthritis-associated interstitial lung disease: clinical characteristics and predictors of mortality. *Respiration* 2019;98:455–60.
41 Sparks JA, He X, Huang J, *et al*. Rheumatoid arthritis disease activity predicting incident clinically apparent rheumatoid arthritis-associated interstitial lung disease: a prospective cohort study. *Arthritis Rheumatol* 2019;71:1472–82.