medRxiv preprint doi: https://doi.org/10.1101/2024.05.08.24306389; this version posted May 8, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Highlights

Occupation Recognition and Exploitation in Rheumatology Clinical Notes: Employing Deep Learning Models for Named Entity Recognition and Knowledge Discovery in Electronic Health Records

Alfredo Madrid-García,Inés Pérez-Sancristóbal,Leticia-Leon,Lydia-Abásolo,Benjamín Fernández-Gutiérrez,Luis Rodríguez-Rodríguez

- Deep learning models hold significant potential for structuring and leveraging information in rheumatology
- Diagnoses related to highly disabling mechanical pathology were associated with a higher probability of occupation collection
- Cleaners, helpers, and social workers occupations are linked to mechanical pathologies such as back pain

medRxiv preprint doi: https://doi.org/10.1101/2024.05.08.24306389; this version posted May 8, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Occupation Recognition and Exploitation in Rheumatology Clinical Notes: Employing Deep Learning Models for Named Entity Recognition and Knowledge Discovery in Electronic Health Records

Alfredo Madrid-García[a,*,1], Inés Pérez-Sancristóbal[a,*,1], Leticia-Leon[a], Lydia-Abásolo[a], Benjamín Fernández-Gutiérrez[a,b] and Luis Rodríguez-Rodríguez[a]

[a]*Grupo de Patología Musculoesquelética. Hospital Clínico San Carlos. Instituto de Investigación Sanitaria San Carlos (IdISSC), Prof. Martin Lagos s/n, Madrid, 28040, Spain*

[b]*Facultad de Medicina, Universidad Complutense de Madrid, Pl. de Ramón y Cajal, s/n, Moncloa - Aravaca, Madrid, 28040, Spain*

ARTICLE INFO

Keywords:
artificial intelligence
transformers
named entity recognition
real-world data
rheumatology
natural language processing
occupations
electronic health record
work disability
occupational medicine

ABSTRACT

Occupation is considered a Social Determinant of Health (SDOH) and its effects have been studied at multiple levels. Although the inclusion of such data in the Electronic Health Record (EHR) is vital for the provision of clinical care, specially in rheumatology where work disability prevention is essential, occupation information is often either not routinely documented or captured in an unstructured manner within conventional EHR systems. Encouraged by recent advances in natural language processing and deep learning models, we propose the use of novel architectures (i.e., transformers) to detect occupation mentions in rheumatology clinical notes of a tertiary hospital, and to whom those occupations belongs. We also aimed to evaluate the clinical and demographic characteristics that influence the collection of this SDOH; and the association between occupation and patients' diagnosis. Bivariate and multivariate logistic regression analysis were conducted for this purpose.

A Spanish pre-trained language model, RoBERTa, fine-tuned with biomedical texts was used to detect occupations. The best model achieved a F1-score of 0.725 identifying occupation mentions. Moreover, highly disabling mechanical pathology diagnoses (i.e., back pain, muscle disorders) were associated with a higher probability of occupation collection. Ultimately, we determined the professions most closely associated with more than ten categories of musculoskeletal disorders.

1. Introduction

The relationship between occupation and health has been studied for more than a century [1]. Nowadays, the benefits of incorporating occupation-related information in the Electronic Health Record (EHR), such as a more accurate diagnosis or better health outcomes, have been widely characterized [2, 3], and promoted by agencies such as the National Institute for Occupational Safety and Health (NIOSH) or European Agency for Safety and Health at Work (EU-OSHA) [4]. Although the relevance of this Social Determinant of Health (SDOH) [5] has been identified and studied at multiple health levels (i.e., mental health [6], health inequality [7], or self-rated health [8]), its study is overshadowed by other SDOH including gender, race, or ethnicity [9, 10]; compromising its content [11], variability [12] and quality [13]. Moreover, most SDOH are stored as free-text unstructured data [14] making them difficult to handle and use.

In the field of Rheumatic and Musculoskeletal Diseases (RMDs), the need to accurately capture occupation information is crucial for the provision of care and to promote prevention and intervention activities that can reduce

This work was supported by the Instituto de Salud Carlos III, Ministry of Health, Madrid, Spain [RD21/002/0001]. The sponsor or funding organization had no role in the design or conduct of this research. The journal's Fee was funded by the institution employing the senior author of the manuscript (Fundación Biomédica del Hospital Clínico San Carlos)

The authors declare there are no competing interests

*Corresponding author

alfredo.madrid.garcia@alumnos.upm.es (A. Madrid-García)

ORCID(s): 0000-0002-1591-0467 (A. Madrid-García); 0000-0002-2098-4313 (I. Pérez-Sancristóbal); 0000-0002-6126-8786 (B. Fernández-Gutiérrez); 0000-0002-2869-7861 (L. Rodríguez-Rodríguez)

[1]First author

medRxiv preprint doi: https://doi.org/10.1101/2024.05.08.24306389; this version posted May 8, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

work disability and sick leave [15]. For instance, recent studies have shown an association between occupation and different inflammatory and non-inflammatory musculoskeletal diseases, such as low back pain [16], Osteoarthritis of the knee (OAk) [17], epicondylitis [18], fibromyalgia severity [19], systemic lupus erythematosus [20], rheumatoid arthritis [21, 22], and presence of anti-citrullinated protein antibodies (ACPA) [23] and anti-neutrophil cytoplasmic antibodies (ANCA) [24].

Automatic detection of information in free-text notes using Natural Language Processing (NLP) techniques and its classification in pre-defined categories is called Named Entity Recognition (NER). This information extraction task has benefited from recent advances in Deep Learning (DL) architectures such as transformers. Briefly, these models are trained on a vast amount of data in an unsupervised way to learn the general structure of a language and its vocabulary [25, 26]. Once trained, the weights of the neurons comprising the model are updated using task-specific data in a process called fine-tuning. In this way, models learn to solve specific tasks in a particular domain. Transformers have been applied in the past in rheumatology for the classification of temporal artery biopsy reports [27], or to identify and classify medical entities in clinical notes related to rheumatoid arthritis [28]. Until today, studies that focus on the patient's occupation are scarce, and even more, studies that attempt to identify occupation mentions in EHR using advanced DL approaches, for knowledge discovery. To address this gap and with the aim of characterizing the patient's occupation, we proposed the use of transformers for occupation NER in a rheumatology departmental EHR. By structuring occupational data, we pursue to characterised the relevance of this SDOH in rheumatology. Therefore, the objective of this study is three-fold:

1. To assess the performance of NER models when extracting occupation mentions in rheumatology clinical notes.
2. To describe the demographic and clinical characteristics that influence the collection of occupation-related information.
3. To analyze the association between occupation and patient's diagnosis.

2. Material and Methods

2.1. Hospital Clínico San Carlos Musculoskeletal Cohort and inclusion criteria

Clinical narratives from the Hospital Clínico San Carlos Musculoskeletal Cohort (HCSC-MSKC) were retrieved. The HCSC-MSKC is a routine clinical practice cohort that includes more than 117,000 visits and 35,000 subjects seen at the rheumatology outpatient clinic from 1st April 2007 to 30th November 2017. More details about this data source can be found elsewhere [29]. The sex-age distribution of patients belonging to this cohort at their first visit is shown in Supplementary Figure 1.

Only visits containing free-text information from patients with free-text information in the first visit were retrieved. After a preliminary analysis and to avoid selection bias, only patients classified as *active* in the *labour situation* structured variable from the EHR and aged between 16 and 65 were included.

2.2. Methodology

2.2.1. Methodology for performance of NER models objective

In a previous study [25], we developed a model for occupation detection, understood as a sequence-labelling NER problem; and another model for the attribution of such occupation to the different stakeholders involved in the healthcare process. Both of them are accessible through Hugging Face [30]. They are RoBERTa-based [31] models, trained with Spanish biomedical texts and fine-tuned using data from MEDDOPROF corpus [32]. MEDDOPROF is a public corpus consisting of 1,844 Spanish clinical case reports with annotations for occupations (i.e., occupations that provide a person with an income or livelihood), working status, and activities (i.e., non-remunerated professions); as well as annotations for to whom the occupation belongs, namely, patient, family member, health professional, or others. Hence, the corpus contained a total of seven classes. More details about this corpus and the model development are shown in Supplementary Material *MEDDOPROF corpus description* and *Model development* sections.

In the present study, 2,000 first visit clinical notes from the HCSC-MSKC were randomly chosen for annotation to build a gold standard, and to evaluate the performance of the previous models in a real-world scenario. These 2,000 notes were annotated (i.e., the occupations and the stakeholders contained in the EHR are manually highlighted using software tools) by two annotators, AMG and IPS, using brat rapid annotation tool (BRAT) and following MEDDOPROF corpus annotation guidelines [33]. No data from the HCSC-MSKC was used for fine-tuning or training the models. The Inter-Annotator Agreement (IAA) F1-score was measured using the *bratiaa* python package, accessible through GitHub [34]. Discrepancies between annotators were resolved by consensus.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.08.24306389; this version posted May 8, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Once the gold standard set was built, inference was made locally. The evaluation metrics used to assess the performance of the models were precision, recall and F1-score [35], microaveraged. The overall values of these metrics were calculated using the MEDDOPROF task evaluation script, also accesible through GitHub [36]. The values for each entity were computed using the sequence labeling evaluation (seqeval) Python library [37]. Confusion matrices at the token level are provided in the Supplementary Material, *Supplementary tables* section.

After analysing the models' performance, predictions were made on the rest of the HCSC-MSKC notes. Overlapping entities (i.e., same text recognised by the two models), classified as "profession-patient"; as well as, entities classified as "profession" (i.e., the text is only recognised by the first model) or "patient" were retrieved, as we found that we could recover approximately 10% of professions attributable to patients, see Supplementary Material *Annotation process and gold standard* and *Predictions in the HCSC-MSKC notes* sections.

2.2.2. Methodology for demographic and clinical characteristics that influence occupation collection objective

Once the predictions were made and manually reviewed, the European Skills, Competencies, Qualifications and Occupations (ESCO) classification version 1.1.1 was used to normalize the patients' occupations. Visits with more than one occupation (e.g., past and present occupation or moonlighter) were also considered. Only the first visit where the patient's occupation was mentioned was obtained, then the nearest neighbor matching method (1:1) was used to identify visits of patients without any occupation mention, with *MatchIt* R package [38], Supplementary Material *Matching visits*. Age, sex, visit number, time since first visit to the rheumatology clinical until current visit and calendar year variables were used to conduct the matching. Logistic regression bivariate and multivariate analyses were conducted to identify clinical and demographic predictors associated with the occupation collection (i.e., occupation collection as the dependent variable, with clinical and demographic factors as independent variables). Firstly, only predictors with a prevalence > 5% in the group of visits with occupation mentions were evaluated, otherwise, they were reclassified into broader categories until the minimum prevalence was reached, see *Supplementary Excel File Predictors Classification*. The predictors used in the matching were not included in the analyses. Secondly, predictors with a p-value < 0.15 in the bivariate analyses were included in the multivariate analyse. On the other hand, the predictors selection in the multivariate analysis was conducted following an hybrid stepwise approach, optimising the Akaike Information Criterion (AIC) value. Variables with a p-value > 0.05 were excluded.

2.2.3. Methodology for association between occupation and patient's diagnosis objective

To assess the relation between professions and diagnoses, we conducted bivariate and multivariate logistic regression analyses, using each disease as the dependent variable and the occupations as independent variables. To this end, ESCO codes were re-classified into categories to ensure a minimum prevalence of 30 mentions per occupation category, see *Supplementary Excel File ESCO*. Only patients visits with occupation mentions were studied. As in the previous objective, predictors (i.e., occupations) with a p-value < 0.15 in the bivariate analyses were included in the multivariate analysis. A multivariate analysis, following an hybrid stepwise approach optimising the AIC, was performed for each disease selected in the previous step with a minimum prevalence of 5%.

Python 3.8.16 was utilized to fine-tune the NER models, and R 4.3.1 was employed in the desriptive and statistical analyses.

2.3. Statistical analysis

Structured demographic and clinical variables, including sex, age, visit type, disease, or Quality of Life (QoL) were used to characterise patients and visits with and without occupation mentions. Dichotomous and categorical variables were summarised using proportions. Continuous variables were summarised using the median and first and third quartiles (Q1–Q3).

2.4. Ethics board approval and reporting guidelines

HCSC Ethics Review Board approval (23/340-E) was obtained as a retrospective study and waiver of informed consent was obtained for the use of unidentified clinical records. Furthermore, the study was conducted in accordance with the Declaration of Helsinki. MI-CLAIM checklist was used to report the results of this study [39], see *Supplementary Material MI-CLAIM checklist*.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.08.24306389; this version posted May 8, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

3. Results

3.1. Performance of NER models in HCSC-MSKC gold standard set

The precision, recall, and the microaveraged F1-score values in the gold standard set, for the occupation detection task, were 0.806, 0.633, and 0.709; for the actors associated with the occupation 0.783, 0.702, and 0.740; and combined 0.795, 0.668, and 0.725. In the gold standard set, only 8.05% of notes (i.e., 161 out of 2,000) contained one "profession-patient" mention and six notes contained more than one "profession-patient" mention. The total number of "profession-patient" mentions was 167, see Supplementary Table 7.

3.2. Demographic and clinical characteristics that influence occupation collection

3.2.1. Predictions in the HCSC-MSKC notes

Predictions were made on a total of 117,068 free-text notes from 35,470 patients. From 5,917 visits in which at least one entity was recognised as "patient", "profession" or "profession-patient", 3,978 visits (67.22%) from 3,723 patients (10.5%) have a real occupation registered, after conducting a manual review. Hence, only the 3.40% out of the whole HCSC-MSKC notes set have occupation-related information. Supplementary Figure 2 shows the percentage of occupation collection stratified by physician.

Considering the demographics and clinical characteristics of the patients, Table 1, and that only 218 out of the 3,723 (5.86%) patients have more than one visit with an occupation mention, we opted to only analysed the visits in which the occupation was collected for the first time, irrespective of whether it was the patient's first visit or not. Finally, the rest of the inclusion criteria were applied; see Figure 1.

3,527 patients out of 3,723 had at least one occupation mention that met the inclusion criteria. Of these, 515 patients with no occupation information at their first visit had that information collected in a subsequent visit. Supplementary Figure 3 shows a histogram of the time elapsed, days, until occupation collection for these 515 patients.

The 3,527 visits with occupation mentions were compared to 3,527 paired visits of patients who did not have any occupation mention visits, after matching. Table 2 shows the differences between both groups.

3.2.2. Bivariate and multivariate analysis

After re-classifying predictors according to their prevalence, 13 disease groups remained. The list of variables assessed in the bivariate analyses can be found in Supplementary Material *Bivariate and multivariates analyses* section.

Table 3 shows the multivariate analysis results. As it can be observed, diagnoses related to highly disabling mechanical conditions (i.e., back pain, muscle disorders, tendinitis) were associated with a higher probability of occupation collection. However, mild distress and slight disability do not appear to have a significant impact on the likelihood of occupational information collection.

3.3. Association between occupation and patient's diagnosis

A total of 402 ESCO codes were used to manually normalise the data. The codes were later re-classified into 34 categories, see Supplementary Excel File ESCO classification. Figure 2 shows the OR of occupations that were significantly associated with each diagnosis and Supplementary Table 12 present the results of the multivariate analysis on the association between the occupation and the patient's diagnosis. As it can be appreciated, "*cleaners and helpers*" and "*social work associte professionals*" exhibit similar patterns: they are associated with mechanical diseases such as back pain, Osteoarthritis of the knee (OAk) or muscle disorders. On one hand, "*electrical and electronic trades workers*", as well as "*sales and purchasing agents and brokers*", are more commonly associated with OAk. On the other hand, "*sports and fitness workers*" and "*garment and related trades workers*" are predominantly associated with other joint diseases. "*Food preparation assistants*" tend to suffer from peripheral neuropathy. "*Customer service clerks*" are exclusively linked to autoimmune pathologies. Finally, "*Hairdressers, beauticians, and related workers*" occupation play a similar behaviour in autoimmune and tendinitis of the upper extremities diseases.

4. Discussion

In this work, we have explored the occupation collection prevalence in a departmental EHR; the sociodemographic characteristics of patients from whom occupation information was gathered compared to those without such data; and the association between occupation and patient's diagnosis. This has been made possible thanks to the use of novelty NLP DL architectures (i.e., transformers) for NER. The use of AI techniques by the rheumatology research community has been described elsewhere [40, 41]. As reported in previous studies, this SDOH is collected in a very low

medRxiv preprint doi: https://doi.org/10.1101/2024.05.08.24306389; this version posted May 8, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

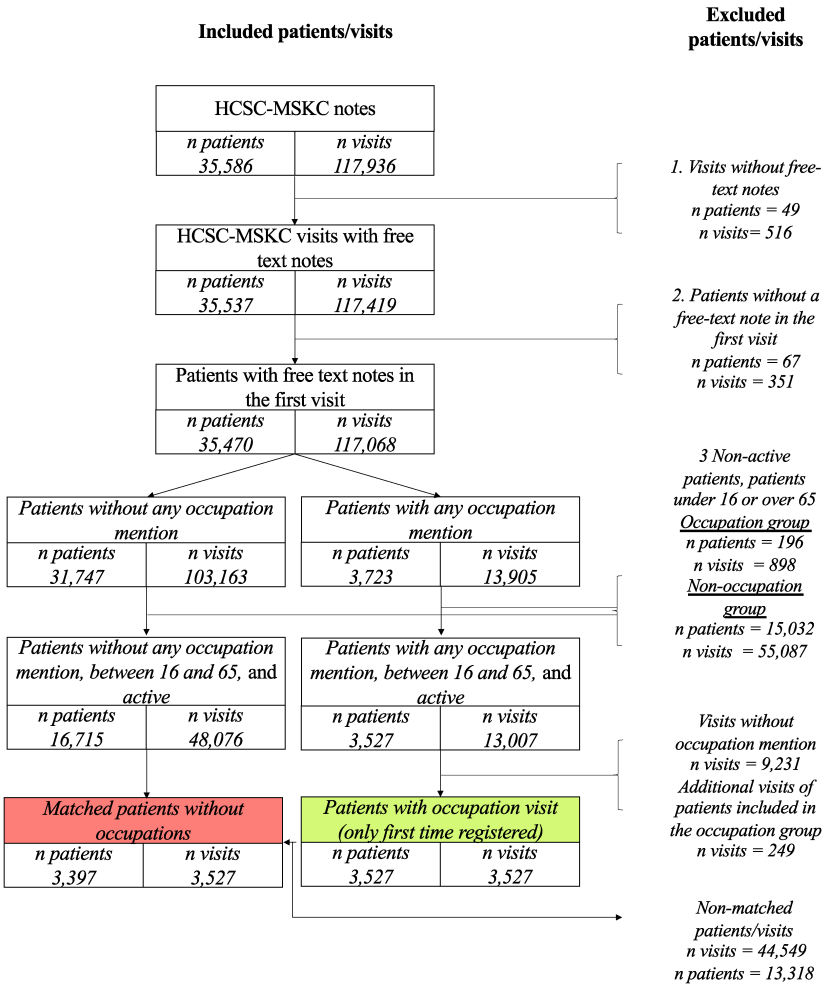

Figure 1: Inclusion criteria diagram

proportion in the EHR. In the biomedical literature, occupation recognition has been recently studied in depth in [25], and this task has been commonly addressed simultaneously with the recognition of other SDOH [42, 43, 44, 45, 46]. A review published in 2021 [10] covered the approaches used for extracting different SDOH. Of 6,402 publications, 82 met the inclusion criteria and only 7 included occupational information. The approaches used in these seven articles consisted of rule-based algorithms. The authors in [47] presented a 10-step method for developing and validating an application to text-mining occupations from psychiatric clinical notes. A noteworthy finding was that the percentage of patients with an occupation recorded increased from 14% to 57% when considering unstructured fields. On the other hand, Dehghan et al [48] developed a large occupation dictionary used to identify occupation mentions on public and non-public clinical datasets from different institutions and countries. Researchers combined rules and ML algorithms for that purpose. However, when using rules, the results are highly dependent on their quality, have difficulties in dealing with negation, uncertainty, and ambiguity, do not scale well with increasing data size, and lack flexibility and generalizability. Moreover, there is always the risk of rules becoming outdated due to the evolving nature of biomedical language. Similarly, ML approaches also have drawbacks; they require the time-consuming creation of handcrafted features and a large amount of labeled data to perform well [49]. To overcome these drawbacks, approaches centred on DL methods have gained relevance. For instance, as part of the 2022 National NLP Clinical Challenges (n2c2) task 2, authors in [50] trained a transformer model with SHAC corpus achieving a 0.85-0.90 F1-score when extracting the employment status. In this work, our F1-score is lower, however it is noteworthy to note that corpus and other NLP resources are not easily accessible in languages other than English [51], and the training data is scarce.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.08.24306389; this version posted May 8, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

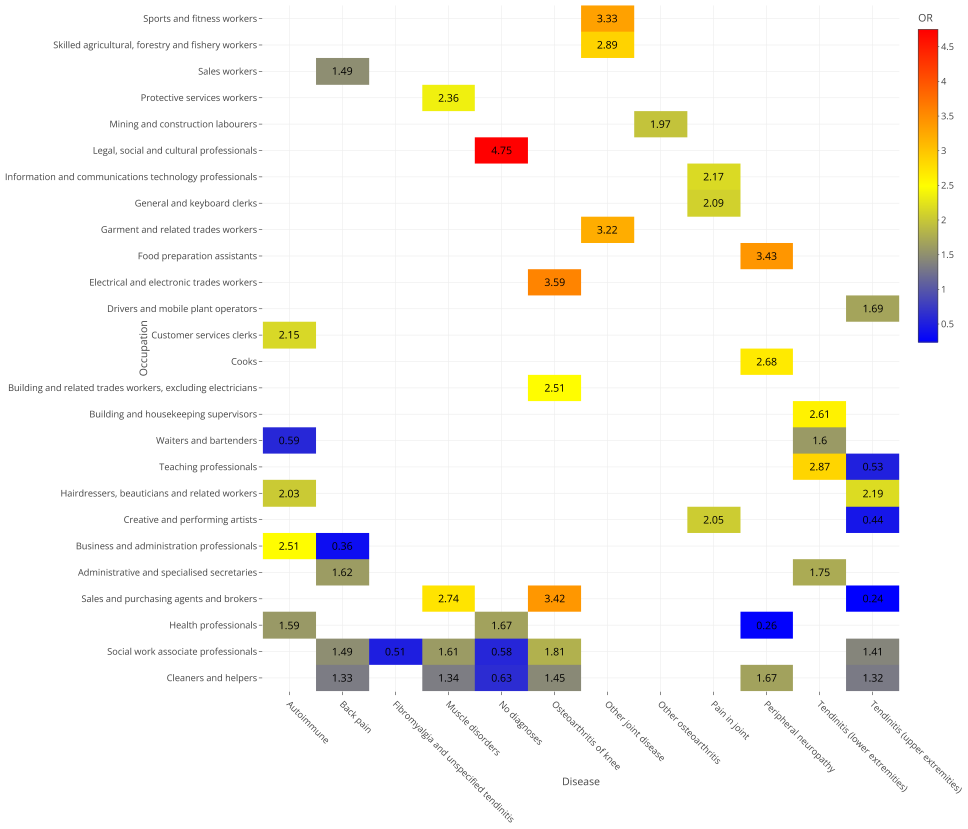

Figure 2: Heatmap showing the OR of occupations that were significantly associated with diagnoses. Neck pain did not contain any associated occupation and therefore was not included. The multivariate analyses results can be found in Supplementary Table 13

As shown in objective 2, the presence of mechanical pathologies promotes the occupation collection. However, this might be influenced by the fact that mechanical disorders are those more related to work disability and sick leave. Rheumatologists from the HCSC have developed and implemented different work disability programmes over the last decades in different health areas of Madrid, which could have enriched the collection of occupation for these pathologies [52].

The relationship between cleaners and mechanical pathologies (e.g., back, neck, shoulders, elbows, hands, and lower limbs), has been reported previously by the EU-OSHA [53]. In [54, 55] authors also described the association between RMDs, such as upper limb disorders, and this profession. Moreover, the association between autoimmune diseases such as rheumatoid arthritis and hairdressers and beauticians was also described more than twenty years ago [56]. It is believed that direct contact with toxic substances may be involved. Readymade garment workers and joint disease relation has also been shown in the past [57]. Futhermore, distal ulnar neuropathy has been reported in chefs [58]. All these findings are aligned with our results. Eventually, there are other significant associations such as OAk and electrical workers or brokers that have not been described and that could be relevant to examine further.

4.1. Strenghts

One of the primary strengths of this study is its innovative and novel approach to detect occupations in rheumatology-related unstructured texts using transformers models. Moreover, to the best of our knowledge there has been no previous study attempting to demonstrate the relationship between occupations and diseases in a rheumatology outpatient electronic medical record. Therefore, this study could serve as a starting point for further exploration of occupation as a SDOH and its implications in rheumatology. Eventually, the findings of this study are consistent with existing literature on the subject.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.08.24306389; this version posted May 8, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

4.2. Limitations

This study has certain technical limitations:

- The NER models were not fine-tuned on data from the HCSC-MSKC. By using data from this cohort, the model could have learned local expressions, syntax nuances or abbreviations that could have resulted in higher performance. The reason why fine-tuning was not performed using HCSC-MSKC data was because a cloud computing environment, not GDPR compliant, was employed for training. Nevertheless, in low-budget settings and in centres without access to specialized hardware, NER models with acceptable performance can be trained using publicly available corpus resources and cloud computing services. However, rheumatology-specific corpus for training AI systems are scarce. Recent efforts, such as RheumaLpack [59], are trying to solve this.
- The MEDDOPROF training set did not contain any rheumatology clinical notes, see Supplementary Table 1. However, as occupation entities are not exclusively collected in a concrete medical speciality, we assume that this did not significantly affect the models performance.
- In this work, two models for two different tasks were trained and their outputs combined. Other authors have experimented with the cross-concatenation of the occupation classes and the family relation classes, training a single model [60]. However, a reduction in performance of 10% was noticed.
- Automatic occupation normalization task to a common terminology was not addressed in this work. Conversely, we manually normalized each recognised entity.
- Not all rheumatologists involved in HCSC-MSKC collect occupations with the same frequency, but the distribution of new patients is random among them (i.e., they will not see a pathology more frequently with respect to the others).

5. Conclusion

We have accurately identified occupation mentions in real-life rheumatology clinical notes using novel NLP approaches. The acquisition of occupation-related information is only collected in a small percentage of patients, around 10%, posing significant challenges to its integration into clinical decision-making processes. Diagnoses related to highly disabling mechanical pathologies were associated with a higher probability of occupation collection by the phyisician. Ultimately, we have shown the association between occupation and more than ten diagnoses. Our findings largely align with existing literature, although we have also identified some unreported associations between occupations and RMDs.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.08.24306389; this version posted May 8, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Table 1
Clinical and sociodemographic characteristics of all visits with occupation mentions

Characteristics	Visits with occupation mentions (n = 3,978)
Women	2,619 (65.84%)
Age	47.67 (38.98, 54.52)
Visit number	
1	3,167 (79.61%)
2	347 (8.72%)
3	125 (3.14%)
4	69 (1.73%)
>4	270 (6.79%)
Time from first visit (days)	570.00 (162.00, 1,155.00)
Calendar year	
2007	99 (2.49%)
2008	161 (4.05%)
2009	253 (6.36%)
2010	317 (7.97%)
2011	379 (9.53%)
2012	415 (10.43%)
2013	489 (12.29%)
2014	449 (11.29%)
2015	452 (11.36%)
2016	509 (12.80%)
2017	455 (11.44%)
Labour situation	
Active	3,808 (95.73%)
Homemaker	30 (0.75%)
Retired	128 (3.22%)
Student	12 (0.30%)
Distress	
No distress	869 (21.85%)
Mild	2,347 (59.00%)
Moderate	741 (18.63%)
Severe	21 (0.53%)
Disability	
1	1,886 (47.41%)
2	1,305 (32.81%)
3	490 (12.32%)
4	243 (6.11%)
5	50 (1.26%)
6	4 (0.10%)
7	0 (0.00%)
Rosser	0.99 (0.98, 0.99)
Training doctor visit	185 (4.65%)

6. Tables

medRxiv preprint doi: https://doi.org/10.1101/2024.05.08.24306389; this version posted May 8, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Table 2
Clinical and sociodemographic characteristics for both groups. ǂ excluding first visits

	Matched visits w/o occupation mentions (N = 3,527)	**Visits with occupation mentions (N = 3,527)**
Women	2,364 (67.03%)	2,331 (66.09%)
Age	46.78 (37.60, 53.90)	46.95 (38.36, 53.36)
Visit number		
1	2,913 (82.59%)	3,012 (85.40%)
2	338 (9.58%)	236 (6.69%)
3	83 (2.35%)	79 (2.24%)
4	50 (1.42%)	45 (1.28%)
>4	143 (4.05%)	155 (4.39%)
Time from first visitǂ (days)	251.00 (44.00, 933.75)	581.00 (155.50, 1,140.50)
Calendar year		
2007	109 (3.09%)	95 (2.69%)
2008	196 (5.56%)	153 (4.34%)
2009	166 (4.71%)	235 (6.66%)
2010	254 (7.20%)	285 (8.08%)
2011	334 (9.47%)	328 (9.30%)
2012	379 (10.75%)	372 (10.55%)
2013	444 (12.59%)	425 (12.05%)
2014	420 (11.91%)	399 (11.31%)
2015	404 (11.45%)	400 (11.34%)
2016	441 (12.50%)	445 (12.62%)
2017	380 (10.77%)	390 (11.06%)
Labour situation		
Active	3,527 (100.00%)	3,527 (100.00%)
Distress		
No distress	687 (19.48%)	791 (22.43%)
Mild	2,275 (64.50%)	2,072 (58.75%)
Moderate and severe	565 (16.02%)	664 (18.83%)
Disability		
No disability	1,593 (45.17%)	1,707 (48.40%)
Slight disability	1,344 (38.11%)	1,121 (31.78%)
Disability	590 (16.73%)	699 (19.82%)
Rosser	0.99 (0.99, 0.99)	0.99 (0.98, 0.99)
Training doctor visit	192 (5.44%)	148 (4.20%)
Disease		
Back pain	469 (13.30%)	842 (23.87%)
Tendinitis (upper extremities)	485 (13.75%)	711 (20.16%)
Pain in Joint	452 (12.82%)	412 (11.68%)
Neck pain	160 (4.54%)	402 (11.40%)
Muscle disorders	107 (3.03%)	382 (10.83%)
No diagnoses	616 (17.47%)	370 (10.49%)
Autoimmune	454 (12.87%)	321 (9.10%)
Other osteoarthritis	278 (7.88%)	298 (8.45%)
Tendinitis (lower extremities)	217 (6.15%)	237 (6.72%)
Knee osteoarthritis	138 (3.91%)	225 (6.38%)
Fibromyalgia and unspecified tendinitis	210 (5.95%)	217 (6.15%)
Peripheral neuropathy	119 (3.37%)	198 (5.61%)
Other joint disease	298 (8.45%)	183 (5.19%)

medRxiv preprint doi: https://doi.org/10.1101/2024.05.08.24306389; this version posted May 8, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Table 3
Multivariate analysis results after conducting hybrid stepwise feature selection. Second objective: association between occupation collection and patients' clinic and demographic characteristics

Variable	OR	p-value
(Intercept)	0.97 (0.85-1.1)	0.642
Back pain	1.94 (1.69-2.23)	2.30E-20
Tendinitis (upper extremities)	1.8 (1.56-2.07)	3.40E-16
Muscle disorders	2.6 (2.05-3.32)	9.30E-15
Peripheral neuropathy	2.06 (1.62-2.63)	4.85E-09
Osteoarthritis of knee	1.84 (1.47-2.31)	1.60E-07
Neck pain	1.64 (1.32-2.03)	6.02E-06
No diagnosis	0.77 (0.66-0.9)	0.001
Other joint disease	0.77 (0.63-0.94)	0.011
Pain in joint	1.16 (0.99-1.36)	0.062
Distress		
No distress	Ref.	
Mild distress	0.72 (0.63-0.82)	1.02E-06
Moderate and severe distress	0.89 (0.74-1.08)	0.254
Disability		
No disability	Ref.	
Slight disability	0.78 (0.7-0.88)	7.20E-05
Disability	1 (0.85-1.18)	0.998

medRxiv preprint doi: https://doi.org/10.1101/2024.05.08.24306389; this version posted May 8, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

A. Appendix

CRediT authorship contribution statement

Alfredo Madrid-García: Conceptualization of this study, methodology, annotation, coding, writing (original draft preparation), feature engineering, review. **Inés Pérez-Sancristóbal:** Annotation, feature engineering. **Leticia-Leon:** Conceptualization of this study, review. **Lydia-Abásolo:** Conceptualization of this study, review. **Benjamín Fernández-Gutiérrez:** Conceptualization of this study. **Luis Rodríguez-Rodríguez:** Conceptualization of this study, methodology, feature engineering, review.

All of the authors were involved in the drafting and/or revising of the manuscript.

Data availability statement The transformers' models used in this study are openly available in Hugging Face `https://huggingface.co/HCSCRheuma/Occupations` [30].

Supplementary material files

- Supplementary Material: expanded methodology section with additional supplementary tables and figures.
- Supplementary Excel File Predictors Classification: classification of clinical and demographic predictors into new categories to conduct logistic regression analysis as stated in objective 2 and 3.
- Supplementary Excel File ESCO Classification: classification of occupations into new categories to conduct logistic regression analysis as stated in objective 3.
- Supplementary material MI-CLAIM checklist

References

[1] S. W. Abbott. Effect of occupation on health. *Publications of the American Statistical Association*, 3(20):238–240, 1892. ISSN 15225437. URL `http://www.jstor.org/stable/2276622`.

[2] David H. Wegman, Catharyn T. Liverman, Andrea M. Schultz, and Larisa M. Strawbridge. *Incorporating occupational information in electronic health records: Letter report*. National Academies Press, 10 2011. ISBN 0309217431. doi: 10.17226/13207.

[3] Matthew Schmitz and Linda Forst. Industry and occupation in the electronic health record: An investigation of the national institute for occupational safety and health industry and occupation computerized coding system. *JMIR Med Inform*, 4(1):e5, 02 2016. ISSN 2291-9694. doi: 10.2196/medinform.4839. URL `https://doi.org/10.2196/medinform.4839`.

[4] National occupational research agenda, April 1996. URL `https://stacks.cdc.gov/view/cdc/21295`. Book.

[5] Jennifer O'Neill, Hilary Tabish, Vivian Welch, Mark Petticrew, Kevin Pottie, Mike Clarke, Tim Evans, Jordi Pardo Pardo, Elizabeth Waters, Howard White, et al. Applying an equity lens to interventions: using progress ensures consideration of socially stratifying factors to illuminate inequities in health. *Journal of clinical epidemiology*, 67(1):56–64, 2014.

[6] Ana Llena-Nozal, Maarten Lindeboom, and France Portrait. The effect of work on mental health: does occupation matter? *Health Economics*, 13(10):1045–1062, 2004. doi: https://doi.org/10.1002/hec.929. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/hec.929`.

[7] Bastian Ravesteijn, Hans van Kippersluis, and Eddy van Doorslaer. *The Contribution of Occupation to Health Inequality*. Emerald Group Publishing Limited, 12 2013. doi: 10.1108/S1049-2585(2013)0000021014.

[8] Ralitza Gueorguieva, Jody L. Sindelar, Tracy A. Falba, Jason M. Fletcher, Patricia Keenan, Ran Wu, and William T. Gallo. The Impact of Occupation on Self-Rated Health: Cross-Sectional and Longitudinal Evidence from the Health and Retirement Survey. *The Journals of Gerontology: Series B*, 64B(1):118–124, 02 2009. ISSN 1079-5014. doi: 10.1093/geronb/gbn006. URL `https://doi.org/10.1093/geronb/gbn006`.

[9] Stacey Marovich, Genevieve Barkocy Luensman, Barbara Wallace, and Eileen Storey. Opportunities at the intersection of work and health: Developing the occupational data for health information model. *Journal of the American Medical Informatics Association*, 27(7):1072–1083, 06 2020. ISSN 1527-974X. doi: 10.1093/jamia/ocaa070. URL `https://doi.org/10.1093/jamia/ocaa070`.

[10] Braja G Patra, Mohit M Sharma, Veer Vekaria, Prakash Adekkanattu, Olga V Patterson, Benjamin Glicksberg, Lauren A Lepow, Euijung Ryu, Joanna M Biernacka, Al'ona Furmanchuk, Thomas J George, William Hogan, Yonghui Wu, Xi Yang, Jiang Bian, Myrna Weissman, Priya Wickramaratne, J John Mann, Mark Olfson, Jr Campion, Thomas R, Mark Weiner, and Jyotishman Pathak. Extracting social determinants of health from electronic health records using natural language processing: a systematic review. *Journal of the American Medical Informatics Association*, 28(12):2716–2727, 10 2021. ISSN 1527-974X. doi: 10.1093/jamia/ocab170. URL `https://doi.org/10.1093/jamia/ocab170`.

[11] Lily A Cook, Jonathan Sachs, and Nicole G Weiskopf. The quality of social determinants data in the electronic health record: a systematic review. *Journal of the American Medical Informatics Association*, 29(1):187–196, 10 2021. ISSN 1527-974X. doi: 10.1093/jamia/ocab199. URL `https://doi.org/10.1093/jamia/ocab199`.

[12] Elizabeth A Lindemann, Elizabeth S Chen, Yan Wang, Steven J Skube, and Genevieve B Melton. Representation of social history factors across age groups: a topic analysis of free-text social documentation. In *AMIA Annual Symposium Proceedings*, volume 2017, page 1169. American Medical Informatics Association, 2017.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.08.24306389; this version posted May 8, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

[13] Ranyah Aldekhyyel, Elizabeth S Chen, Sripriya Rajamani, Yan Wang, and Genevieve B Melton. Content and quality of free-text occupation documentation in the electronic health record. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2016:1708–1716, 2016. ISSN 1942-597X.

[14] Elham Hatef, Gurmehar Singh Deol, Masoud Rouhizadeh, Ashley Li, Katyusha Eibensteiner, Craig B. Monsen, Roman Bratslaver, Margaret Senese, and Hadi Kharrazi. Measuring the value of a practical text mining approach to identify patients with housing issues in the free-text notes in electronic health record: Findings of a retrospective cohort study. *Frontiers in Public Health*, 9, 2021. ISSN 2296-2565. doi: 10.3389/fpubh.2021.697501. URL https://www.frontiersin.org/articles/10.3389/fpubh.2021.697501.

[15] William S. Marras, Robert G. Cutlip, Susan E. Burt, and Thomas R. Waters. National occupational research agenda (nora) future directions in occupational musculoskeletal disorder health research. *Applied Ergonomics*, 40(1):15–22, 2009. ISSN 0003-6870. doi: https://doi.org/10.1016/j.apergo.2008.01.018. URL https://www.sciencedirect.com/science/article/pii/S0003687008000343.

[16] Sara E Luckhaupt, James M Dahlhamer, Gabriella T Gonzales, Ming-Lun Lu, Matthew Groenewold, Marie Haring Sweeney, and Brian W Ward. Prevalence, recognition of work-relatedness, and effect on work of low back pain among u.s. workers. *Annals of Internal Medicine*, 171 (4):301–304, 2019. doi: 10.7326/M18-3602. URL https://www.acpjournals.org/doi/abs/10.7326/M18-3602. PMID: 31083729.

[17] Xia Wang, Thomas A. Perry, Nigel Arden, Lingxiao Chen, Camille M. Parsons, Cyrus Cooper, Lucy Gates, and David J. Hunter. Occupational risk in knee osteoarthritis: A systematic review and meta-analysis of observational studies. *Arthritis Care & Research*, 72(9):1213–1223, 2020. doi: https://doi.org/10.1002/acr.24333. URL https://acrjournals.onlinelibrary.wiley.com/doi/abs/10.1002/acr.24333.

[18] Karen Walker-Bone, Keith T. Palmer, Isabel Reading, David Coggon, and Cyrus Cooper. Occupation and epicondylitis: a population-based study. *Rheumatology*, 51(2):305–310, 10 2011. ISSN 1462-0324. doi: 10.1093/rheumatology/ker228. URL https://doi.org/10.1093/rheumatology/ker228.

[19] A B Mohabbat, E C Wight, N M L Mohabbat, S Nanda, J A Ferguson, L Philpot, and J Adusumalli. The correlation between occupation type and fibromyalgia severity. *Occupational Medicine*, 73(5):257–262, 05 2023. ISSN 0962-7480. doi: 10.1093/occmed/kqad063. URL https://doi.org/10.1093/occmed/kqad063.

[20] Christine G. Parks and Glinda S. Cooper. Occupational exposures and risk of systemic lupus erythematosus. *Autoimmunity*, 38(7):497–506, 2005. doi: 10.1080/08916930500285493. URL https://doi.org/10.1080/08916930500285493. PMID: 16373255.

[21] Giovanni Adami, Marco Pontalti, Giorgio Cattani, Maurizio Rossini, Ombretta Viapiana, Giovanni Orsolini, Camilla Benini, Eugenia Bertoldo, Elena Fracassi, Davide Gatti, and Angelo Fassio. Association between long-term exposure to air pollution and immune-mediated diseases: a population-based cohort study. *RMD Open*, 8(1), 2022. doi: 10.1136/rmdopen-2021-002055. URL https://rmdopen.bmj.com/content/8/1/e002055.

[22] Bowen Tang, Qianwen Liu, Anna Ilar, Pernilla Wiebert, Sara Hägg, Leonid Padyukov, Lars Klareskog, Lars Alfredsson, and Xia Jiang. Occupational inhalable agents constitute major risk factors for rheumatoid arthritis, particularly in the context of genetic predisposition and smoking. *Annals of the Rheumatic Diseases*, 82(3):316–323, 2023. ISSN 0003-4967. doi: 10.1136/ard-2022-223134. URL https://ard.bmj.com/content/82/3/316.

[23] Anna Ilar, Lars Alfredsson, Pernilla Wiebert, Lars Klareskog, and Camilla Bengtsson. Occupation and risk of developing rheumatoid arthritis: Results from a population-based case–control study. *Arthritis Care & Research*, 70(4):499–509, 2018. doi: https://doi.org/10.1002/acr.23321. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/acr.23321.

[24] Séverine Beaudreuil, Gérard Lasfargues, Laurence Lauériere, Zen El Ghoul, Florence Fourquet, Cédric Longuet, Jean-Michel Halimi, Hubert Nivet, and Matthias Büchler. Occupational exposure in anca-positive patients: A case-control study. *Kidney International*, 67(5):1961–1966, 2005. ISSN 0085-2538. doi: https://doi.org/10.1111/j.1523-1755.2005.00295.x. URL https://www.sciencedirect.com/science/article/pii/S0085253815506753.

[25] Alfredo Madrid-García. Recognition of professions in medical documentation. Master's thesis, Universidad Nacional de Educación a Distancia (España). Escuela Técnica Superior de Ingeniería Informática. Departamento de Inteligencia Artificial, 2023.

[26] Marco Guevara, Shan Chen, Spencer Thomas, Tafadzwa L Chaunzwa, Idalid Franco, Benjamin H Kann, Shalini Moningi, Jack M Qian, Madeleine Goldstein, Susan Harper, et al. Large language models to identify social determinants of health in electronic health records. *npj Digital Medicine*, 7(1):1–14, 2024.

[27] Christopher McMaster, Victor Yang, Benjamin Sutu, Shereen Oon, Gene-Siew Ngian, Ian Wicks, Russell Buchanan, and David Liew. Temporal artery biopsy reports can be accurately classified by artificial intelligence. In *ARTHRITIS & RHEUMATOLOGY*, volume 73, pages 2962–2964. WILEY 111 RIVER ST, HOBOKEN 07030-5774, NJ USA, 2021.

[28] Meiting Li, Feifei Liu, Jia'an Zhu, Ran Zhang, Yi Qin, and Dongping Gao. Model-based clinical note entity recognition for rheumatoid arthritis using bidirectional encoder representation from transformers. *Quantitative Imaging in Medicine and Surgery*, 12(1), 2021. ISSN 2223-4306. URL https://qims.amegroups.org/article/view/77073.

[29] Alfredo Madrid-García, Judit Font-Urgelles, Mario Vega-Barbas, Leticia León-Mateos, Dalifer Dayanira Freites, Cristina Jesus Lajas, Esperanza Pato, Juan Angel Jover, Benjamín Fernández-Gutiérrez, Lydia Abásolo-Alcazar, and Luis Rodríguez-Rodríguez. Outpatient readmission in rheumatology: A machine learning predictive model of patient's return to the clinic. *Journal of Clinical Medicine*, 8(8), 2019. ISSN 2077-0383. doi: 10.3390/jcm8081156. URL https://www.mdpi.com/2077-0383/8/8/1156.

[30] Alfredo Madrid García. Occupations (revision 92bdacc), 2023. URL https://huggingface.co/HCSCRheuma/Occupations.

[31] Casimiro Pio Carrino, Joan Llop, Marc Pàmies, Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Joaquín Silveira-Ocampo, Alfonso Valencia, Aitor Gonzalez-Agirre, and Marta Villegas. Pretrained biomedical language models for clinical NLP in Spanish. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 193–199, Dublin, Ireland, 06 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.bionlp-1.19. URL https://aclanthology.org/2022.bionlp-1.19.

[32] Salvador Lima-López, Eulàlia Farré-Maduell, Antonio Miranda-Escalada, Vicent Briva-Iglesias, and Martin Krallinger. MEDDOPROF corpus: complete gold standard annotations for occupation detection in medical documents in Spanish, 03 2022. URL https://doi.org/10.5281/zenodo.7116201.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.08.24306389; this version posted May 8, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

[33] Eulàlia Farré-Maduell, Salvador Lima-López, Antonio Miranda-Escalada, Vicent Briva-Iglesias, and Martin Krallinger. Meddoprof guidelines, April 2021. URL https://doi.org/10.5281/zenodo.4720833.

[34] Tobias Kolditz. Bratiaa. https://github.com/kldtz/bratiaa, 2019.

[35] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147, 2003. URL https://aclanthology.org/W03-0419.

[36] TeMU-BSC. Meddoprof evaluation library, 2021. URL https://github.com/TeMU-BSC/meddoprof-evaluation-library. Software available from https://github.com/TeMU-BSC/meddoprof-evaluation-library.

[37] Hiroki Nakayama. seqeval: A python framework for sequence labeling evaluation, 2018. URL https://github.com/chakki-works/seqeval. Software available from https://github.com/chakki-works/seqeval.

[38] Daniel E. Ho, Kosuke Imai, Gary King, and Elizabeth A. Stuart. MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, 42(8):1–28, 2011. doi: 10.18637/jss.v042.i08.

[39] Beau Norgeot, Giorgio Quer, Brett K Beaulieu-Jones, Ali Torkamani, Raquel Dias, Milena Gianfrancesco, Rima Arnaout, Isaac S Kohane, Suchi Saria, Eric Topol, et al. Minimum information about clinical artificial intelligence modeling: the mi-claim checklist. *Nature medicine*, 26(9):1320–1324, 2020.

[40] Alfredo Madrid-García, Beatriz Merino-Barbancho, Alejandro Rodríguez-González, Benjamín Fernández-Gutiérrez, Luis Rodríguez-Rodríguez, and Ernestina Menasalvas-Ruiz. Understanding the role and adoption of artificial intelligence techniques in rheumatology research: an in-depth review of the literature. *Seminars in Arthritis and Rheumatism*, page 152213, 2023. ISSN 0049-0172. doi: https://doi.org/10.1016/j.semarthrit.2023.152213. URL https://www.sciencedirect.com/science/article/pii/S0049017223000550.

[41] Alfredo Madrid-García, Zulema Rosales-Rosado, Dalifer Freites-Nuñez, Inés Pérez-Sancristóbal, Esperanza Pato-Cour, Chamaida Plasencia-Rodríguez, Luis Cabeza-Osorio, Lydia Abasolo-Alcázar, Leticia León-Mateos, Benjamín Fernández-Gutiérrez, and Luis Rodríguez-Rodríguez. Harnessing chatgpt and gpt-4 for evaluating the rheumatology questions of the spanish access exam to specialized medical training. *Scientific Reports*, 13:22129, 12 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-49483-6.

[42] Kevin Lybarger, Mari Ostendorf, and Meliha Yetisgen. Annotating social determinants of health using active learning, and characterizing determinants using neural event extraction. *Journal of Biomedical Informatics*, 113:103631, 2021. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2020.103631. URL https://www.sciencedirect.com/science/article/pii/S1532046420302598.

[43] Rachel Stemerman, Jaime Arguello, Jane Brice, Ashok Krishnamurthy, Mary Houston, and Rebecca Kitzmiller. Identification of social determinants of health using multi-label classification of electronic health record clinical notes. *JAMIA Open*, 4(3), 02 2021. ISSN 2574-2531. doi: 10.1093/jamiaopen/ooaa069. URL https://doi.org/10.1093/jamiaopen/ooaa069. ooaa069.

[44] Sifei Han, Robert F. Zhang, Lingyun Shi, Russell Richie, Haixia Liu, Andrew Tseng, Wei Quan, Neal Ryan, David Brent, and Fuchiang R. Tsui. Classifying social determinants of health from unstructured electronic health records using deep learning-based natural language processing. *Journal of Biomedical Informatics*, 127:103984, 2022. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2021.103984. URL https://www.sciencedirect.com/science/article/pii/S1532046421003130.

[45] Shaina Raza, Elham Dolatabadi, Nancy Ondrusek, Laura Rosella, and Brian Schwartz. Discovering social determinants of health from case reports using natural language processing: Algorithmic development and validation. *medRxiv*, 2022. doi: 10.1101/2022.11.30.22282946. URL https://www.medrxiv.org/content/early/2022/12/06/2022.11.30.22282946.

[46] Zehao Yu, Xi Yang, Chong Dang, Prakash Adekkanattu, Braja Gopal Patra, Yifan Peng, Jyotishman Pathak, Debbie L. Wilson, Ching-Yuan Chang, Wei-Hsuan Lo-Ciganic, Thomas J. George, William R. Hogan, Yi Guo, Jiang Bian, and Yonghui Wu. Soda: A natural language processing package to extract social determinants of health for cancer studies, 2022. URL https://arxiv.org/abs/2212.03000.

[47] Natasha Chilman, Xingyi Song, Angus Roberts, Esther Tolani, Robert Stewart, Zoe Chui, Karen Birnie, Lisa Harber-Aschan, Billy Gazard, David Chandran, Jyoti Sanyal, Stephani Hatch, Anna Kolliakou, and Jayati Das-Munshi. Text mining occupations from the mental health electronic health record: a natural language processing approach using records from the clinical record interactive search (cris) platform in south london, uk. *BMJ Open*, 11(3), 2021. ISSN 2044-6055. doi: 10.1136/bmjopen-2020-042274. URL https://bmjopen.bmj.com/content/11/3/e042274.

[48] Azad Dehghan, Tom Liptrot, Daniel Tibble, Matthew Barker-Hewitt, and Goran Nenadic. Identification of occupation mentions in clinical narratives. In Elisabeth Métais, Farid Meziane, Mohamad Saraee, Vijayan Sugumaran, and Sunil Vadera, editors, *Natural Language Processing and Information Systems*, pages 359–365, Cham, 2016. Springer International Publishing. ISBN 978-3-319-41754-7.

[49] Oswaldo Solarte Pabón, Orlando Montenegro, Maria Torrente, Alejandro Rodríguez González, Mariano Provencio, and Ernestina Menasalvas. Negation and uncertainty detection in clinical texts written in spanish: a deep learning-based approach. *PeerJ Computer Science*, 8:e913, 3 2022. ISSN 2376-5992. doi: 10.7717/peerj-cs.913.

[50] Russell Richie, Victor M Ruiz, Sifei Han, Lingyun Shi, and Fuchiang (Rich) Tsui. Extracting social determinants of health events with transformer-based multitask, multilabel named entity recognition. *Journal of the American Medical Informatics Association*, 04 2023. ISSN 1527-974X. doi: 10.1093/jamia/ocad046. URL https://doi.org/10.1093/jamia/ocad046. ocad046.

[51] Aurélie Névéol, Hercules Dalianis, Sumithra Velupillai, Guergana Savova, and Pierre Zweigenbaum. Clinical natural language processing in languages other than english: opportunities and challenges. *Journal of Biomedical Semantics*, 9:12, 12 2018. ISSN 2041-1480. doi: 10.1186/s13326-018-0179-8.

[52] Lydia Abásolo, Margarita Blanco, Javier Bachiller, Gloria Candelas, Paz Collado, Cristina Lajas, Marcelino Revenga, Patricia Ricci, Pablo Lázaro, Maria Dolores Aguilar, et al. A health system program to reduce work disability related to musculoskeletal disorders. *Annals of Internal Medicine*, 143(6):404–414, 2005. doi: 10.7326/0003-4819-143-6-200509200-00005. URL https://www.acpjournals.org/doi/abs/10.7326/0003-4819-143-6-200509200-00005. PMID: 16172439.

[53] European Agency for Safety and Health at Work. E-fact 39 - cleaners and musculoskeletal disorders. https://osha.europa.eu/en/publications/e-fact-39-cleaners-and-musculoskeletal-disorders, 2017. "Accessed: 2024-02-02".

medRxiv preprint doi: https://doi.org/10.1101/2024.05.08.24306389; this version posted May 8, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

[54] Haimanot Melese, Tsiwaye Gebreyesus, Abayneh Alamer, and Awtachew Berhe. Prevalence and associated factors of musculoskeletal disorders among cleaners working at mekelle university, ethiopia. *Journal of Pain Research*, pages 2239–2246, 2020.

[55] Mei Ching Lim, Khamisah Awang Lukman, Nelbon Giloi, Jac Fang Lim, Richard Avoi, Syed Sharizman Syed Abdul Rahim, and Mohammad Saffree Jeffree. Prevalence of upper limb musculoskeletal disorders and its associated risk factors among janitorial workers: A cross-sectional study. *Annals of Medicine and Surgery*, 73:103201, 2022.

[56] Åsa Reckner Olsson, Thomas Skogh, and Gun Wingren. Occupational determinants for rheumatoid arthritis. *Scandinavian journal of work, environment & health*, pages 243–249, 2000.

[57] Humayun Kabir, Myfanwy Maple, Kim Usher, and Md Shahidul Islam. Health vulnerabilities of readymade garment (rmg) workers: a systematic review. *BMC Public Health*, 19:1–20, 2019.

[58] Roberta Bonfiglioli, Stefano Mattioli, and Francesco S Violante. Occupational mononeuropathies in industry. *Handbook of clinical neurology*, 131:411–426, 2015.

[59] Alfredo Madrid, Beatriz Merino Barbancho, Dalifer Dayanira Freites Nuñez, Luis Rodriguez Rodriguez, Ernestina Menasalvas Ruiz, Alejandro Rodriguez Gonzalez, and Anselmo Peñas. From web to rheumalpack: Creating a linguistic corpus for artificial intelligence exploitation and knowledge discovery in rheumatology. *medRxiv*, 2024. doi: 10.1101/2024.04.26.24306269. URL https://www.medrxiv.org/content/early/2024/04/27/2024.04.26.24306269.

[60] Salvador Medina Herrera and Jorge Turmo Borras. Everything transformers: Recognition, classification and normalisation of professions and family relations. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021): co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021), XXXVII International Conference of the Spanish Society for Natural Language Processing: Málaga, Spain, September, 2021*, pages 770–775. CEUR-WS. org, 2021.