**Supplementary materials:**

**Incorporating natural language processing to improve classification of axial spondyloarthritis using electronic health records**

Sizheng Steven Zhao, Chuan Hong, Tianrun Cai, Chang Xu, Jie Huang, Joerg Ermann, Nicola J Goodson, Daniel H Solomon, Tianxi Cai, Katherine P Liao

**Methods**

Disease definition

Since EHR data were primarily recorded for patient care rather than research, investigations required to assess fulfilment of research classification criteria might be omitted when patients attend with a previously established diagnosis. We additionally categorised patients with incomplete data for the modified New York criteria as AS if they had imaging pathology consistent with AS on plain-film radiographs, CT or MRI (e.g. ankylosis of the vertebrae or sacroiliac joint fusion). Patients with incomplete data for the ASAS criteria were categorised as nr-axSpA if they had suggestive changes on CT or MRI (e.g. sacroiliac joint sclerosis/irregularity). Our prior work showed that patients with complete and pragmatic fulfilment of classification criteria were similar, with no difference in gender, SpA features (e.g., HLA-B27 status, family history, uveitis, psoriasis) or disease severity (e.g., ESR/CRP and pain).

Surrogate labels

Three surrogate labels were created to select the most informative features from those identified using online knowledge resources. Patients were labeled as cases or non-cases if they: 1) had <1 or >10 AS ICD codes, 2) had <1 or >10 AS NLP concepts, or 3) met a known criterion with high PPV (≥3 ICD-9 or 10 codes for AS and string-text mention of “sacroiliitis”, “ankylosis” or “syndesmophyte” (whether present or absent) in radiology reports, including plain X-rays, MRIs or CT scans).

No / not available

550 randomly selected patients form the axSpA Mart

≥ Grade 2 sacroiliitis on sacroiliac joint (SIJ) radiograph, or clinical notes stating the same

82 definite AS

Yes

Definitive radiographic pathology of the vertebrae (e.g. ankylosis, syndesmophytes); or definitive pathology of the SIJ or vertebrae (e.g. fusion or ankylosis) on CT/MRI; or clinical notes stating the same

8 AS partially fulfilling the modified NY criteria

Bone marrow oedema in SIJ or vertebrae on MRI; or fulfils ASAS clinical criteria

21 definite nr-axSpA

Yes

Bony changes seen on CT/MRI (e.g. SIJ sclerosis/irregularity or syndesmophytes); or clinical notes stating mild or early radiographic sacroiliitis, without further description or grading

Clinical notes stating mild/early radiographic sacroiliitis, without further description or grading

16 nr-axSpA partially fulfilling the ASAS criteria

Diagnosis of AS but no evidence on radiology reports or clinical notes

62 axSpA with physician diagnosis but no available evidence for classification criteria

No documented diagnosis of AS; has an alternative diagnosis; or do not fit into other categories

No / not available

No / not available

No / not available

No / not available

361 not axSpA

Yes

Yes

Supplementary figure 1. Classification process for 550 randomly selected patients for the gold-standard training set. Severity of sacroiliitis was determined using a combination of radiology reports and scoring by a trained rheumatologist (SSZ). Both definite and partial classification criteria axSpA (in green) were defined as cases for the training set, while cases with insufficient/no evidence (in red) were not. ASAS: Assessment of Spondyloarthritis International Society; NY: New York; SIJ: sacroiliac joint; nr-axSpA: non-radiographic axial SpA.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Supplementary table 1. Patient characteristics of the enriched cohort and distributions of their codified and narrative data. | | | | |
|  | | Classified as axSpA cases | Not classified as axSpA cases | P |
| N | | 1509 | 6344 |  |
| Age, years | | 53 (17) | 57 (18) | <0.001 |
| Male | | 1070 (71%) | 3067 (48%) | <0.001 |
| Race | White | 1199 (79%) | 5167 (81%) | <0.001 |
| African American | 52 (3%) | 270 (4%) |
| Other | 258 (17%) | 907 (14%) |
| ICD codes for AS | | 6 (2, 14) | 0 (0, 1) | <0.001 |
| NLP for AS | | 25 (10, 56) | 1 (0, 3) | <0.001 |
| NLP for spondylitis | | 3 (0, 8) | 0 (0, 0) | <0.001 |
| NLP for HLA-B27 | | 0 (0, 4) | 0 (0, 1) | <0.001 |
| NLP for sacroiliitis | | 1 (0, 4) | 0 (0, 1) | <0.001 |
| Healthcare utilization | | 47 (18, 111) | 84 (30, 201) | <0.001 |
| Data shown as mean (sd), number (percentage) or median (interquartile range). | | | | |