RHEUMATOLOGY

Rheumatology 2022;61:581–590
doi:10.1093/rheumatology/keab310
Advance Access publication 26 March 2021

Original article

Using Bayesian networks to identify musculoskeletal symptoms influencing the risk of developing psoriatic arthritis in people with psoriasis

Amelia Green [1], William Tillett[1,2], Neil McHugh[1,2,*] and Theresa Smith[3,*]; on behalf of the PROMPT Study Group

Abstract

Objectives. The aim of this study was to explore the use of Bayesian networks (BNs) to understand the relationships between musculoskeletal symptoms and the development of PsA in people with psoriasis.

Methods. Incident cases of psoriasis were identified for 1998 to 2015 from the UK Clinical Research Practice Datalink. Musculoskeletal symptoms (identified by Medcodes) were concatenated into primary groups, each made up of several subgroups. Baseline demographics for gender, age, BMI, psoriasis severity, alcohol use and smoking status were also extracted. Several BN structures were composed using a combination of expert knowledge and data-oriented modelling based on: (i) primary musculoskeletal symptom groups; (ii) musculoskeletal symptom subgroups and (iii) demographic variables. Predictive ability of the networks using the area under the receiver operating characteristic curve was calculated.

Results. Over one million musculoskeletal symptoms were extracted for the 90 189 incident cases of psoriasis identified, of which 1409 developed PsA. The BN analysis yielded direct relationships between gender, BMI, arthralgia, finger pain, fatigue, hand pain, hip pain, knee pain, swelling, back pain, myalgia and PsA. The best BN, achieved by using the more site-specific musculoskeletal symptom subgroups, was 76% accurate in predicting the development of PsA in a test set and had an area under the receiver operating characteristic curve of 0.73 (95% CI: 0.70, 0.75).

Conclusion. The presented BN model may be a useful method to identify clusters of symptoms that predict the development of PsA with reasonable accuracy. Using a BN approach, we have shown that there are several symptoms which are predecessors of PsA, including fatigue, specific types of pain and swelling.

Key words: psoriatic arthritis, psoriasis, musculoskeletal symptoms, Bayesian network, Clinical Practice Research Datalink

CLINICAL SCIENCE

Rheumatology key messages

- Limited information on the preclinical phase of PsA exists.
- Our study suggests there are several musculoskeletal symptoms that are predecessors of PsA.
- Patients who exhibit these symptoms may have PsA earlier than their PsA diagnosis.

[1]Department of Pharmacy and Pharmacology, University of Bath, [2]Royal National Hospital for Rheumatic Diseases, NHS Foundation Trust and [3]Department of Mathematical Sciences, University of Bath, Bath, UK

Submitted 21 October 2020; accepted 12 March 2021

*Neil McHugh and Theresa Smith contributed equally to this study.

Correspondence to: Amelia Green, Department of Pharmacy and Pharmacology, University of Bath, Claverton Down, Bath, Avon, BA2 7AY, UK. E-mail: aj409@bath.ac.uk

Introduction

Psoriasis is a common skin disorder that affects over one million people in England [1]. Approximately 20% of people with psoriasis are also affected by a progressive and often destructive joint disease known as PsA [2]. PsA is preceded by psoriasis in ~70% of cases, and although there is a clinical overlap between the two diseases, the inflammation, joint

damage, and deformity associated with PsA adds substantially to the disease burden and can result in severe disability and reduced quality of life [3]. Despite this major impact, the clinical features influencing the risk of PsA in people with psoriasis are poorly understood. With evidence now supporting early intervention in PsA, investigating the presence and characteristics of the preclinical phase of PsA (the period before musculoskeletal inflammation becomes clinically detectable) is of great interest [4]. Understanding the pre-diagnosis period could be key in the earlier identification of PsA. As PsA may not develop until years after the onset of psoriasis [5], patients with psoriasis represent a unique population for identifying symptoms related to the development of PsA, especially in the early phases of the disease.

Studies investigating the preclinical phases of PsA are limited owing to difficulties related to the heterogeneous nature of PsA and the subtlety of the findings from physical examination. In addition, the overlap of PsA with OA and mechanical/overuse soft tissue disorders in many older psoriasis patients can cause issues when trying to identify the disease. Around 50% of the patients with psoriasis who are seen in primary care settings have persistent musculoskeletal symptoms [6]. While, in the majority of cases, these symptoms could be attributed to other non-inflammatory conditions, given the heterogeneous nature of PsA and the difficulties in identifying some of its clinical features, it is possible that some of these patients have preclinical PsA. Such was the case in a study by Eder *et al.*, who found that the presence of a complex of non-specific musculoskeletal symptoms, including joint pain, fatigue and stiffness, in patients with psoriasis was representative of a preclinical phase of PsA in the absence of objective findings of the disease [7].

In addition to the challenge of distinguishing patients with true preclinical PsA from psoriasis patients whose symptoms are related to other non-inflammatory rheumatic conditions, characterizing the relationships between symptoms is made difficult by their highly dependent nature, which violates the (conditional) independence assumption of most models. Approaches that move away from single-response models (i.e. regression analyses) and allow one to investigate relations between several PsA-related symptoms simultaneously are therefore needed in order to make inferences about the relationships present. Bayesian networks (BNs), a class of probabilistic graphical models, are naturally suited to representing a set of variables and their conditional dependencies, in addition to being able to handle uncertainty. That being the case, in this article we investigate their use in order to generate models that can be used to identify and characterize the relationships between musculoskeletal symptoms and the development of PsA in people with psoriasis.

Methods

Data source

This study used data from the Clinical Practice Research Datalink (CPRD), a UK electronic healthcare database that contains anonymized longitudinal medical records for >20 million patients [8, 9]. The patients contributing data to the CPRD are generally representative of the UK general population in terms of age, sex and ethnicity.Data were captured using structured hierarchical classifications called Read codes and Medcodes [10].

Study population

This study used a cohort that we have previously described [11]. Briefly, the base population consisted of all incident cases of psoriasis, aged 16–89 years at the time of diagnosis, who were permanently registered and contributing to the CPRD at any time during the study period, with at least 1 year of valid data collection [8]. Cases of psoriasis were predominantly identified based on Read codes, which have been demonstrated to have a high validity [12], but psoriasis-specific treatment information was also used as an indication of a psoriasis diagnosis. Cases were only classified as incident if they had a minimum of 1 year of valid data collection before their date of diagnosis [13]. Incident cases of PsA were identified from within the base population based on Read codes, which have been found to have a high positive predictive value in a similar UK database [14]. The date of PsA diagnosis was taken as the date of the first PsA code or date of first DMARD prescription in the absence of evidence of an alternative indication for the DMARD, whichever was earlier. Patients with a diagnosis of PsA on the same day as or prior to their psoriasis diagnosis date were excluded.

Selection of model variables

Baseline demographics for gender, age, BMI, alcohol use, smoking status, and psoriasis severity were extracted from the CPRD using methods that we have previously described [11]. BMI was classified based on the World Health Organization BMI categories: <25.0, 25.0–29.9, 30.0–34.9, ≥35.0. Due to a high degree of missing data (>50%), baseline BMI measurements were estimated using a patient's most recent BMI value in the 3 years before their psoriasis index date. Smoking status was classified as smoker, ex-smoker or non-smoker. Alcohol consumption status was categorized as heavy drinker (>6 U/day), drinker, ex-drinker or non-drinker. Psoriasis was classified as severe if patients had prescriptions for medicines consistent with the treatment of severe disease (systemic therapies) or evidence of a referral to a dermatologist; it was classified as mild in the absence of evidence of severe disease. For those patients who developed PsA during the follow-up period, psoriasis severity was classified using data before the date of PsA diagnosis.

Musculoskeletal symptoms occurring during the study period were identified based on Medcodes (a numeric equivalent of the Read code used within the CPRD). All distinct Medcodes across all patient histories were extracted. All Medcodes were assessed by rheumatologists (N.J.M. and W.T.) for suitably of inclusion. Redundant Medcodes (Medcodes that were unlikely to represent a musculoskeletal symptom associated with the development of PsA), along with those relating to other non-inflammatory rheumatic conditions, were discarded. Due to the high volume of remaining distinct Medcodes (300+ Medcodes) and the fact that multiple Medcodes can represent the same musculoskeletal symptom, Medcodes were concatenated into several groups. These groupings were based on findings from similar studies and clinical knowledge from rheumatologists [6, 7]. Six primary groups were identified: arthritis, bursitis or enthesitis, fatigue, pain, stiffness, or swelling. Each primary group was broken out into a subset of smaller subgroups that were generally more site-specific (45 subgroups in total). The Medcode compositions of these 6 primary groups and 45 subgroups are shown in Supplementary Table S1, available at *Rheumatology* online. Although some patients had multiple, identical symptom Medcodes recorded within the study period, binary features were created for all groups. While BNs can handle discrete and continuous data, the creation of binary rather than count features is thought to reduce the effect of frequency of general practitioner (GP) visits in the data [15].

Bayesian network analysis and statistical methods

BNs were used to identify and characterize the relationships between baseline demographic variables, musculoskeletal symptoms, and PsA. BNs are based on a graphical formalism called a Directed Acyclic Graph. In a BN, each variable is modelled as a node, and the relationships between these variables are represented by edges. Directed edges (arcs) go from a *parent* node to a *child* node. In this modelling, we do not assume a causal interpretation of the arcs in the network, primarily because BNs themselves are not inherently causal models, but also because the design of our study does not allow us to confirm any causal relationships. Instead, we interpret the arcs as direct dependence relationships between the linked variables, and the absence of arcs means the existence of conditional independence relationships.

Several BN models were constructed. To begin, a BN model was learnt using the six primary musculoskeletal symptom groups (BN-1). Then, one-by one, each primary group was broken out into its subgroups and a subsequent BN network was constructed. For each BN, we utilized the Markov Blanket of the PsA node (the set of nodes that have influence on its conditional distribution, i.e. the parents and children of the PsA node, and the parents of the children of the PsA node) as the criterion to select the relevant features [16]. The variables identified as influencing the development of PsA from each of these BNs were then used to build the final BN (BN-2). For comparison, a BN was constructed using only the demographic variables, i.e. no musculoskeletal symptoms, (BN-3).

The data were split at random into 80% for training and 20% for testing. Using the training dataset, the structure of each of the BNs was learnt using a combination of statistical methods and expert knowledge. To begin, we used structural expectation maximization to learn the BN structure from missing data [17]. The expectation maximization algorithm is an iterative method for finding the maximum likelihood estimate of the parameters when the dataset has missing (or hidden) values [18, 19]. Several constraints on edge orientations were placed, including restricting the direction of the arcs for gender and age (i.e. gender and age cannot be influenced by other variables). Afterwards, the validity of each of the BN structures was assessed visually by comparing it with known/expected relationships. Network parameters were learnt using Bayesian estimation, which calculates the expected value of the posterior distribution over the parameters. A Dirichlet prior with an equivalent sample size of one was used as a weak prior over the parameters [16]. In order to assess the predictive ability of the network (i.e. the ability to predict PsA development given a patients demographics and history of musculoskeletal symptoms), the accuracy (defined as the sum of true positive and true negative instances divided by the total number of instances), sensitivity, specificity and the area under the receiver operating characteristic curve were calculated using the test dataset.

Graphical representation of the BN structures

The graphical representations of the BN structures were attained using the 'dot' layout method, which can be attributed to Sugiyama *et al.* [20], offered by the Rgraphviz package [21]. The design goal of the dot layout method is to make aesthetically pleasing drawings of modest-sized graphs. As such, the figures only depict the relationships between the variables, and thus the order/importance of symptoms cannot be inferred from them.

Data were analysed in R version 3.5.0, using the packages bnlearn, pROC, ROCR, ggplot, Rgraphviz and epiR. Further details of the BN analysis can be found in the Supplementary data, available at *Rheumatology* online.

Ethical approval

Ethical approval has been obtained by the CPRD data provider from a Multicentre Research Ethics Committee for all observational studies, and the study protocol was approved by the CPRD Independent Scientific Advisory Committee (15_154R).

TABLE 1 Baseline characteristics of the study population, including the prevalence of musculoskeletal symptom on psoriasis index date

Characteristic	All (n = 90 189)	Psoriasis only (n = 88 780)	PsA (n = 1409)
Sex[a], no. (%) male	43 599 (48)	42 854 (48)	745 (53)
Age[a], mean (S.D.), years	48.3 (18.2)	48.5 (18.2)	44.7 (13.9)
Psoriasis duration,[b] mean (S.D.), years	5.7 (4.1)	5.8 (4.1)	3.6 (3.4)
Psoriasis severity,[a] n (%) severe	11 491 (13)	11 083 (12)	408 (29)
Musculoskeletal symptom,[c] n (%)	4571 (5)	4404 (5)	167 (12)
Arthritis	109 (0)	92 (0)	17 (1)
Bursitis, enthesitis or tendinitis	465 (1)	451 (1)	14 (1)
Fatigue	454 (1)	447 (1)	7 (0)
Pain	3443 (4)	3322 (4)	121 (9)
Stiffness	23 (0)	19 (0)	4 (0)
Swelling	160 (0)	145 (0)	15 (1)
Musculoskeletal duration,[d] mean (S.D.)	4.5 (3.5)	4.5 (3.5)	3.5 (3.1)
BMI category,[e] n (%) (kg/m^2)			
<25	15 519 (17)	15 384 (17)	135 (10)
25.0–29.9	15 841 (18)	15 603 (18)	238 (17)
30.0–34.9	8980 (10)	8818 (10)	162 (11)
≥35	5854 (6)	5728 (6)	126 (9)
Missing	43 995 (49)	43 247 (49)	748 (53)
Smoking status,[e] n (%)			
Non-smoker	38 452 (43)	37 808 (43)	644 (46)
Ex-smoker	25 039 (28)	24 646 (28)	393 (28)
Current smoker	25 133 (28)	24 768 (28)	365 (26)
Missing	1564 (2)	1558 (2)	7 (0)
Alcohol status,[e] n (%)			
Non-drinker	9883 (11)	9745 (11)	138 (10)
Ex-drinker	4430 (5)	4370 (5)	60 (4)
Current-drinker	63 997 (71)	62 924 (71)	1073 (76)
Missing	11 879 (13)	11 741 (13)	138 (10)

[a]On psoriasis index date. [b]Calculated as the period between psoriasis diagnosis and the earliest of (i) the individual developing PsA, (ii) the individual or their practice ceasing to contribute data to the CPRD or (iii) the end of the study period. [c]≥1 musculoskeletal symptom on psoriasis index date. [d]Calculated as the period between psoriasis diagnosis to symptom report date. [e]Closest to and within 3 years prior to psoriasis index date.

Results

Study population

Of the 90 189 patients making up the study population, 1409 developed PsA. The baseline patient characteristics of the study population are summarized in Table 1. The mean (S.D.) age at psoriasis index was 48 years (18), and the mean (S.D.) duration of psoriasis was 5.71 years (4.10).The baseline prevalence of musculoskeletal symptoms was 2.39 times higher in the PsA group (psoriasis patients who went on to develop PsA during the study period) than the psoriasis-only group (psoriasis patients who did not develop PsA during the study period) (95% CI: 2.09, 2.67).

Musculoskeletal symptoms

Over 800 unique musculoskeletal symptom Medcodes were identified from the study population's CPRD records. After review, 327 were deemed suitable for inclusion and were concatenated into the primary groups and their subgroups (Table 2).

Within the study period, a total of 147 647 musculoskeletal symptoms were recorded, with over half of the study population having at least one musculoskeletal symptom recorded during the study period (54 803, 61%). The incidence of musculoskeletal symptoms was significantly higher in the PsA group than in the psoriasis-only group (Table 3 and Supplementary Table S2, available at *Rheumatology* online). Over one-fifth of the psoriasis patients who went on to develop PsA visited their GP with musculoskeletal-related symptoms during the 5 years prior to their diagnosis of PsA. This proportion gradually increased and reached over 57% in the 6 months immediately preceding the diagnosis.

Patients with musculoskeletal symptoms had a longer study duration than those with no musculoskeletal symptoms: mean duration (years) 7.07 *vs* 3.73 psoriasis

Table 2 Grouping of Medcodes into musculoskeletal symptom groups

Primary musculoskeletal symptom groups (*n*, Medcodes)	Musculoskeletal symptom subgroups (*n*, Medcodes)
Arthritis (35)	Ankle arthritis (1), Elbow arthritis (1), Foot arthritis (1), Hand arthritis (5), Hip arthritis (2), Knee arthritis (1), OA (7), Other arthritis (16), Wrist arthritis (1)
Bursitis, Enthesitis or Tendinitis (73)	Bursitis (15), Enthesitis (35), Tendinitis (23)
Fatigue (23)	Fatigue (22), Malaise (1)
Pain (135)	Ankle pain (6), Arm pain (4), Arthralgia (13), Back pain (28), Chest pain (1), Elbow pain (5), FM (5), Finger pain (5), Foot pain (9), Hand pain (6), Hip pain (5), Jaw pain (1), Knee pain (5), Leg pain (5), Myalgia (8), Neck pain (1), Sacroiliitis (1), Sciatica (3), Shoulder pain (10), Unspecified pain (10), Wrist pain (4)
Stiffness (22)	Lower body stiffness (7), Unspecified stiffness (8), Upper body stiffness (7)
Swelling (39)	Ankle swelling (4), Foot swelling (2), Hand swelling (5), Joint swelling (2), Knee swelling (6), Shoulder swelling (3), Unspecified swelling (17)

Table 3 Incidence of musculoskeletal symptoms

Musculoskeletal symptom	Cases	Person years	Incidence rate per 1000 person years (95% CI)
All			
Psoriasis	53 711	243 231	220.82 (218.96, 222.69)
PsA	1092	2104	519.00 (488.22, 549.79)
Arthritis			
Psoriasis	1828	501 928	3.64 (3.47, 3.81)
PsA	188	4797	39.19 (33.59, 44.79)
Bursitis, enthesitis or tendinitis			
Psoriasis	10 119	465 585	21.73 (21.31, 22.16)
PsA	204	4567	44.67 (38.54, 50.8)
Fatigue			
Psoriasis	11 420	459 804	24.84 (24.38, 25.29)
PsA	127	4566	27.81 (22.98, 32.65)
Pain			
Psoriasis	45 374	290 260	156.32 (154.88, 157.76)
PsA	847	2759	307.04 (286.36, 327.72)
Stiffness			
Psoriasis	800	506 921	1.58 (1.47, 1.69)
PsA	41	4961	8.26 (5.73, 10.79)
Swelling			
Psoriasis	6603	482 902	13.67 (13.34, 14.00)
PsA	207	4624	44.77 (38.67, 50.86)

only and 4.14 *vs* 1.51 PsA. While the median number of musculoskeletal symptoms was two for both the psoriasis-only and PsA groups, there were fewer individuals with recorded musculoskeletal symptoms in the psoriasis-only group than in the PsA group (60% psoriasis only *vs* 78% PsA).

When concatenated into the primary groups, the majority of Medcodes were assigned to the pain group (55%). The PsA group were more likely to have an arthritis-, pain- or swelling-related Medcode (15% *vs* 2%, 69% *vs* 55% and 16% *vs* 8%, respectively). Arthritis-related Medcodes were more likely to be non-site-specific in those who developed PsA compared with the psoriasis-only group (14% *vs* 2%).

Bayesian networks

Of the baseline demographic variables considered, psoriasis severity, BMI category, and gender were identified as direct predecessors of PsA, while smoking, alcohol and age influenced PsA through their respective child nodes. Two of the 6 primary musculoskeletal symptom groups (arthritis and swelling), and 10 of the 45 musculoskeletal symptom subgroups (arthralgia, back pain, fatigue, finger pain, hand pain, hip pain, knee pain, myalgia, non-specific arthritis, unspecified swelling) were identified as direct predecessors of PsA.

The results of the model performance measures for the three learned networks, with respect to the test set, can be seen in Table 4. The best area under the

Table 4 Model performance measures for the different learned networks, with respect to the test set

Bayesian network	Accuracy	Sensitivity	Specificity	Area under the receiver operating characteristic curve
BN-1	0.65 (0.64, 0.65)	0.65 (0.64, 0.65)	0.74 (0.69, 0.79)	0.69 (0.67, 0.72)
BN-22	0.76 (0.75, 0.76)	0.76 (0.75, 0.77)	0.70 (0.64, 0.74)	0.73 (0.70, 0.75)
BN-33	0.62 (0.61, 0.63)	0.62 (0.61, 0.62)	0.70 (0.65, 0.75)	0.66 (0.63, 0.68)

BN-1: BN consisting of baseline demographics and the six primary musculoskeletal symptom groups (arthritis, bursitis or enthesitis, fatigue, pain, stiffness, and swelling). BN-2: BN consisting of baseline demographics and only those musculoskeletal symptom subgroups identified as influencing the development of PsA (arthralgia, back pain, fatigue, finger pain, hand pain, hip pain, knee pain, myalgia, non-specific arthritis, unspecified swelling). BN-3: BN consisting of only baseline demographics (i.e. no musculoskeletal symptoms).

receiver operating characteristic curve was 0.73 (95% CI: 0.70, 0.75), translating into 76% sensitivity and 70% specificity, which was achieved by BN-2, the BN using the more site-specific musculoskeletal symptom subgroups. This BN was 76% accurate in predicting the development of PsA. The graphical representation of this BN structure is displayed in Fig. 1 and shows the widespread probabilistic associations between the variables included in the modelling. The graphical representations of the BN structures learnt using (i) the six primary musculoskeletal symptom groups and (ii) only the demographic variables, can be found in the (Supplementary Figs S1 and S2, available at *Rheumatology* online, respectively).

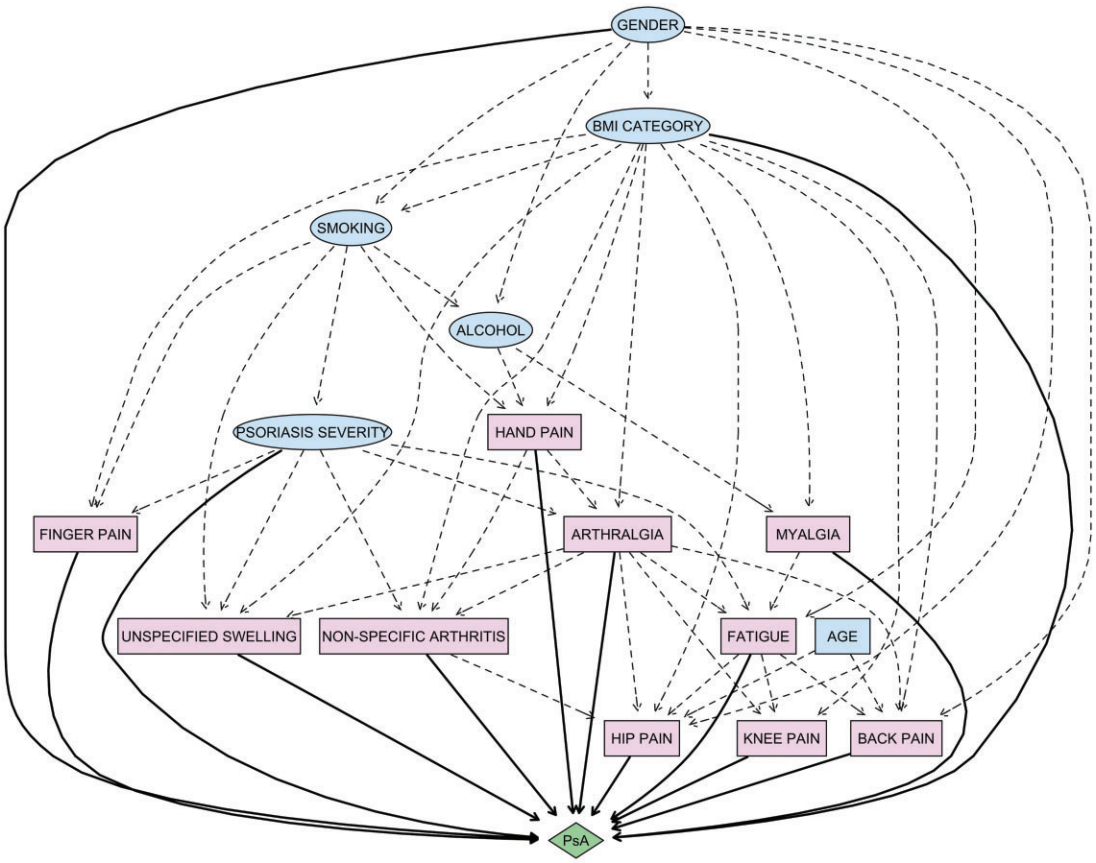
Discussion

As the majority of people destined to develop PsA have a prior presence of psoriasis and a median interval between diagnoses of 7–8 years [5], this unique group provide a valuable opportunity for the early detection of PsA. While individuals with psoriasis and persistent musculoskeletal symptoms may have PsA [6], investigating the relationships between symptoms is challenging, and appropriate statistical approaches are needed in order to make inferences about the relationships present. This article presents the use of a BN approach for discovering the relationships between demographic and musculoskeletal symptom variables, and the development PsA in people with psoriasis. BNs have been used for various applications in a wide variety of domains, including, but not limited to, medical diagnosis, disease prediction, clinical decision-making, and risk prediction [22–27]. BNs not only provide a robust and flexible approach, but they are also able to handle uncertainty and integrate clinical knowledge alongside data-driven methods to infer structure from data. Furthermore, BNs are interpretable, with the relationships of probable factors relating to diseases available in a visual format that is easy to interpret, which is of particular use to those in the healthcare domain (i.e. clinicians/rheumatologists). To the best of our knowledge, this is the first application of a BN analysis to any psoriasis cohort, and the results of our study suggest that there are several musculoskeletal symptoms that are predecessors of PsA. Furthermore, including these musculoskeletal symptoms in our modelling improved the predictive ability of the BN when compared with a BN constructed without musculoskeletal symptom variables. However, we acknowledge that incorporation of additional data (such as blood test results and use of medication) may improve our current model to a stage that can be more readily applied in clinical practice.

Of interest, non-specific arthritis, (a recording of arthritis with no specific type or site specified) was identified as a predecessor to the development of PsA, suggesting that there are many cases in which the presence of arthritis in the context of psoriasis does not immediately lead to a diagnosis of PsA. Similarly, non-specific swelling was identified as a predecessor to the development of PsA, suggesting that disease presentation may be ambiguous and diagnosis challenging. In line with recent findings, fatigue was identified as a predecessor to the development of PsA and therefore should be recognized as an important indicator in future algorithms [28–30]. On the other hand, pain (which is a domain commonly taken into consideration by rheumatologists when assessing a patient with suspected PsA), when considered as a non-specific variable, was not found to influence the development of PsA. However, when pain was broken down into smaller, more specific descriptors, arthralgia, myalgia and pain in the back, finger, hand and knee were all found to have a direct influence on the development of PsA.

In addition to being able to identify the variables that had a direct impact on the development of PsA, the BN analysis was also able to identify how the variables impacted one another. For example, smoking was found to influence psoriasis severity, and thereby PsA. While symptoms such as stiffness, enthesitis and tendinitis, which are common hallmarks of PsA, were not found to influence the development of PsA, it is plausible that patients who did not develop PsA within the study period, who exhibited these symptoms, had undiagnosed PsA. Whether this was due to diagnostic uncertainty, a misdiagnosis, or some other reason remains to be determined. Furthermore, while no constraints were placed on the edge orientations of the PsA node when

Fig. 1 Graphical model of the Bayesian network built using Clinical Practice Research Datalink data to predict the development of PsA in people with psoriasis and to gain insight into the psoriasis population at increased risk of PsA. The figure includes the structure of musculoskeletal symptoms (coloured pink), grouped into categories based on Medcodes related to similar symptoms and defined in terms of counts over the study period, in addition to baseline demographic variables (coloured blue). Nodes represent input variables and edges represent conditional dependencies between the variables. Continuous and discrete variables are indicated by elliptical and rectangular nodes, respectively. Direct and indirect relationships to PsA are indicated solid and dotted black edges (lines). Note, this figure only depicts the relationships between the variables; hence, the order/importance of symptoms cannot be inferred from it



constructing our network, PsA was not found to influence any of the variables included in the modelling.

This BN framework has a clear edge in interpretability over other methods, which is a core requirement for models in medicine, because both patients and clinicians need to understand the results. This framework is also well suited to uncovering and representing relationships between variables, which is key when variables have a highly correlated nature, such as is the case with PsA, and indeed many other rheumatic diseases. Consequently, this framework is not only suited to handle the intricacies of PsA risk prediction, but it could readily be applied within the rheumatology community to other clinical studies using electronic healthcare data, where the interest lies in representing uncertain relationships among important features, disease prediction, and aiding clinical decision-making.

While our study has many strengths, including its population-based nature, the large number of psoriasis patients, the use of validated codes to identify psoriasis and PsA, and the access to original medical records providing more complete information with regards to a patient's history of musculoskeletal symptoms (such as body parts affected), we acknowledge several important limitations. First, it is not possible to obtain a direct confirmation of a rheumatologist's diagnosis of PsA from a primary care record. But, given the difficulty diagnosing PsA, even among rheumatologists, it would be very unlikely for a GP to record a diagnosis of PsA without having referred that patient to a rheumatologist first. Second, it is possible that some GPs may have been more likely to refer a patient with a history of musculoskeletal symptoms to rheumatology for diagnosis of PsA sooner than a patient with no history of musculoskeletal

symptoms. However, by conducting this study in a population of incident cases of psoriasis, we hope to have minimized any impact of diagnostic or surveillance bias on the likelihood of diagnosis of PsA within this population. Third, while all Medcodes representing musculoskeletal symptoms were assessed by rheumatologists experienced in assessing PsA patients, the Medcodes themselves rely upon the complex coding systems and thesauruses built into the CPRD and thus determine what can and will be recorded. Furthermore, recording behaviours of GPs will vary, and detailed information about the joint sites or body parts affected may not always be captured, leading to less precise, less complete and less correct data. Fourth, those with musculoskeletal symptoms recorded during the study period had a longer study duration than those with no musculoskeletal symptoms recorded, suggesting that the presence of a musculoskeletal symptom may, in part, be due to patient follow-up time, and also that some of the PsA patients may have had PsA earlier than the diagnosis date recorded. Fifth, we acknowledge that the data reflect not only the health of the patients, but also patients' interactions with the healthcare system, and thus it is often the case that there is a high degree of missing covariate and response data. Those with non-missing data may be unrepresentative of the general population, and restriction to those with complete data may result in biased analyses [31]. While missing data can affect the learning of the BN structure from data, we hope to have reduced any bias when learning the model structure and parameters in the presence of any missing data through the use of the expectation maximization algorithm alongside clinical domain knowledge and published literature [32]. Finally, no attempt was made to take the timing of symptoms into account in this work, and binary features were used in order to reduce the effect of frequency of GP visits in the data [15]. As a result, information on the intensity of symptoms may have been lost.

In conclusion, we have introduced a BN approach to investigate the relationships between demographic and musculoskeletal symptom variables, and the development PsA in people with psoriasis. The results of our study suggest that there are several musculoskeletal symptoms that are predecessors of PsA, including non-specific arthritis, fatigue, specific types of pain, and swelling. As such, patients who exhibit these symptoms may have PsA earlier than their PsA diagnosis and thus should be followed up more closely, since they are more likely to develop the disease in subsequent years. Further research is still required in order to better understand this preclinical phase of PsA and to improve the identification of psoriasis patients who will go on to develop PsA. In addition, to these clinical findings, we have shown that BNs are able to provide a coherent modelling framework for charaterizing the relationships between musculoskeletal symptoms and the development of PsA in people with psoriasis. In the future, we plan to build on our analysis by including some additional clinical information (i.e. data on tests, prescriptions and referrals). Additionally, in order to assess the intensity and timing of symptoms, we also plan to extend and refine our model using dynamic BNs.

Acknowledgements

We wish to acknowledge the non-contributing authors of the PROMPT (early detection to imPRove OutcoMe in people with undiagnosed Psoriatic arthriTis) study group who have been responsible for the acquisition of funding and general supervision of the research group: Sarah Hewlett, Helen Harris, Philip Helliwell, Laura Coates, Catherine Fernandez, Sarah Brown, Claire Davies, Rachel Charlton, Gavin Shaddick, Alison Nightingale, Julia Snowball, Jonathan Packham, Laura Bjoke, Eldon Spakman, Anne Barton, Oliver Fitzgerald, Vishnu Madhok, Melanie Brooke, Jana James and Andrew Parkinson. This manuscript reports on work previously presented at the ACR Annual meeting 2019; https://acrabstracts.org/abstract/learning-the-relationships-between-psoriatic-arthritis-and-a-patients-history-of-musculoskeletal-symptoms-from-electronic-health-records-using-bayesian-networks/. N.M. contributed to the conception of the study. All authors contributed to the design of the work. Data acquisition and analysis was carried out by A.G. and T.S. All authors were involved in the interpretation of the study results as well as the drafting and revision of the manuscript and all approved the final version to be published. The lead author affirms that this manuscript is an honest, accurate and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

Funding: This report is independent research funded by the National Institute for Health Research (Programme Grants for Applied Research, Early detection to improve outcome in patients with undiagnosed psoriatic arthritis, RP-PG-1212–20007). The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health.

Disclosure statement: All authors have completed the International Committee of Medical Journal Editors uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: all authors report grants from the National Institute for Health Research (RP-PG-1212–20007) during the conduct of the study.

Data availability statement

The data underlying this manuscript were provided by the CPRD. The data are provided on licence from the CPRD (cprd.com) and, as such, we are not able to share the raw data from the study. Read Code lists used to identify the study population have been included as Supplementary data. Medcode lists used to identify

musculoskeletal symptoms will be shared on reasonable request to the corresponding author.

Supplementary data

Supplementary data are available at Rheumatology online.

References

1 Menter A, Korman NJ, Elmets CA *et al.* Guidelines of care for the management of psoriasis and psoriatic arthritis. Section 3. Guidelines of care for the management and treatment of psoriasis with topical therapies. J Am Acad Dermatol 2009;60:643–59.

2 Ogdie A, Weiss P. The epidemiology of psoriatic arthritis. Rheum Dis Clin North Am 2015;41:545–68.

3 Rosen CF, Mussani F, Chandran V *et al.* Patients with psoriatic arthritis have worse quality of life than those with psoriasis alone. Rheumatology (Oxford) 2012;51:571–6.

4 Raychaudhuri SP, Wilken R, Sukhov AC, Raychaudhuri SK, Maverakis E. Management of psoriatic arthritis: early diagnosis, monitoring of disease severity and cutting edge therapies. J Autoimmun 2017;76:21–37.

5 Charlton RA, Tillett W, Nightingale AL *et al.*; the PROMPT Study Group. Interval between onset of psoriasis and psoriatic arthritis: comparing the United Kingdom Clinical Practice Research Datalink with a hospital-based cohort. Rheumatology 2017;56:66.

6 Karreman MC, Weel AEAM, van der Ven M *et al.* Prevalence of psoriatic arthritis in primary care patients with psoriasis. Arthritis Rheumatol 2016;68:924–31.

7 Eder L, Polachek A, Rosen CF *et al.* The development of psoriatic arthritis in patients with psoriasis is preceded by a period of nonspecific musculoskeletal symptoms: a prospective cohort study. Arthritis Rheumatol2017;69:622–9.

8 Herrett E, Gallagher AM, Bhaskaran K *et al.* Data resource profile: Clinical Practice Research Datalink (CPRD). Int J Epidemiol 2015;44:827–36.

9 CPRD. Clinical Practice Research Datalink. http://www.cprd.com (21 August 2020, date last accessed).

10 Chisholm J. The Read clinical classification. Brit Med J 1990;300:1092.

11 Green A, Shaddick G, Charlton R *et al.*; on behalf of the PROMPT study group. Modifiable risk factors and the development of psoriatic arthritis in people with psoriasis. Brit J Dermatol 2020;182:714–20.

12 Seminara NM, Abuabara K, Shin DB *et al.* Validity of The Health Improvement Network (THIN) for the study of psoriasis. Br J Dermatol 2011;164:602–9.

13 Lewis JD, Bilker WB, Weinstein RB, Strom BL. The relationship between time since registration and measured incidence rates in the General Practice Research Database. Pharmacoepidemiol Drug Saf 2005;14:443–51.

14 Ogdie A, Alehashemi S, Love TJ *et al.* Validity of psoriatic arthritis and capture of disease modifying antirheumatic drugs in the health improvement network. Pharmacoepidemiol Drug Saf 2014;23:918–22.

15 Jammeh EA, Carroll CB, Pearson SW *et al.* Machine-learning based identification of undiagnosed dementia in primary care: a feasibility study. BJGP Open 2018;2: bjgpopen18X101589.

16 Koller D, Friedman N. Probabilistic graphical models: principles and techniques. Cambridge, MA: MIT Press, 2009.

17 Friedman N. Learning belief networks in the presence of missing values and hidden variables. In: Proceedings of the Fourteenth International Conference on Machine Learning, vol. 97, 1997. pp. 125–33. Morgan Kaufmann Publishers, San Francisco, CA, USA.

18 Lauritzen SL. The algorithm for graphical association models with missing data. Comput Stat Data Analysis 1995;19:191–201.

19 Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the algorithm. J Royal Stat Soc 1977;39:1–22.

20 Sugiyama K, Tagawa S, Toda M. Methods for visual understanding of hierarchical system structures. IEEE Trans Syst, Man Cybernetics 1981;11:109–25.

21 Ellson J, Gansner ER, Koutsofios E, North SC, Woodhull G. Graphviz and Dynagraph – Static and Dynamic Graph Drawing Tools. https://www.graphviz.org/Documentation/EGKNW03.pdf. (01 July 2020, date last accessed)

22 Farooq K, Hussain A, Leslie S *et al.* (eds). An ontology driven and bayesian network based cardiovascular decision support framework. Berlin, Heidelberg: Springer, 2012.

23 Xie J, Liu Y, Zeng X, Zhang W, Mei Z. A Bayesian network model for predicting type 2 diabetes risk based on electronic health records. Modern Phys Lett B 2017; 31:1740055.

24 Seixas FL, Zadrozny B, Laks J, Conci A, Muchaluat Saade DC. A Bayesian network decision model for supporting the diagnosis of dementia, Alzheimer's disease and mild cognitive impairment. Comput Biol Med 2014;51:140–58.

25 Sambo F, Facchinetti A, Hakaste L *et al.* A Bayesian network for probabilistic reasoning and imputation of missing risk factors in type 2 diabetes. In: Holmes J, Bellazzi R, Sacchi L, Peek N, eds. Artificial intelligence in medicine. AIME 2015. Lecture Notes in Computer Science, vol 9105. Cham: Springer International Publishing, 2015.

26 Loghmanpour NA, Kanwar MK, Druzdzel MJ *et al.* A new Bayesian network-based risk stratification model for prediction of short-term and long-term LVAD mortality. ASAIO J 2015;61:313–23.

27 Fuster-Parra P, Tauler P, Bennasar-Veny M *et al.* Bayesian network modeling: a case study of an epidemiologic system analysis of cardiovascular risk. Comput Methods Programs Biomed 2016;126:128–42.

28 Gudu T, Etcheto A, de Wit M *et al.* Fatigue in psoriatic arthritis – a cross-sectional study of 246 patients from 13 countries. Joint Bone Spine 2016;83:439–43.

29 Eder L, Thavaneswaran A, Chandran V, Cook R, Gladman DD. Factors explaining the discrepancy between physician and patient global assessment of joint and skin disease activity in psoriatic arthritis patients. Arthritis Care Res 2015;67:264–72.

30 Orbai A-M, de Wit M, Mease P *et al.* International patient and physician consensus on a psoriatic arthritis core outcome set for clinical trials. Ann Rheum Dis 2017;76: 673–80.

31 Booth HP, Prevost AT, Gulliford MC. Severity of obesity and management of hypertension, hypercholesterolaemia and smoking in primary care: population-based cohort study. J Hum Hypertens 2016; 30:40–5.

32 Lin J-H, Haug PJ. Exploiting missing clinical data in Bayesian network modeling for predicting medical problems. J Biomed Informatics 2008;41: 1–14.