



Citation for published version:

PROMPT Study Group 2022, 'Using Bayesian Networks to identify musculoskeletal symptoms influencing the risk of developing Psoriatic Arthritis in people with psoriasis', *Rheumatology (Oxford, England)*, vol. 61, no. 2, pp. 581-590. <https://doi.org/10.1093/rheumatology/keab310>

DOI:

[10.1093/rheumatology/keab310](https://doi.org/10.1093/rheumatology/keab310)

Publication date:

2022

Document Version

Peer reviewed version

[Link to publication](#)

This is a pre-copyedited, author-produced version of an article accepted for publication in *Rheumatology* following peer review. The version of record Amelia Green, William Tillet, Neil McHugh, Theresa Smith, the PROMPT study group, Using Bayesian Networks to identify musculoskeletal symptoms influencing the risk of developing Psoriatic Arthritis in people with psoriasis, *Rheumatology*, 2021;, keab310 is available online at: <https://academic.oup.com/rheumatology/advance-article/doi/10.1093/rheumatology/keab310/6189038>

University of Bath

Alternative formats

If you require this document in an alternative format, please contact:
openaccess@bath.ac.uk

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Using Bayesian Networks to identify musculoskeletal symptoms influencing the risk of developing Psoriatic Arthritis in people with psoriasis

Amelia Green¹, William Tillett^{1,2}, *Neil McHugh^{1,2}, *Theresa Smith³, on behalf of the PROMPT study group

¹Department of Pharmacy and Pharmacology, University of Bath, Bath, UK

²Royal National Hospital for Rheumatic Diseases, Upper Borough Walls, Bath, UK

³Department of Mathematical Sciences, University of Bath, Bath, UK

*Joint senior authors

Corresponding Author

Amelia Green

Department of Pharmacy and Pharmacology, University of Bath, Claverton Down, Bath, Avon,
BA2 7AY

Email: aj409@bath.ac.uk

Keywords: Psoriatic arthritis, Psoriasis, Musculoskeletal symptoms, Bayesian network, Clinical Practice Research Datalink

ABSTRACT

Objectives To explore Bayesian networks (BNs) in understanding the relationships between musculoskeletal symptoms and the development of PsA in people with psoriasis.

Methods Incident cases of psoriasis were identified between 1998 and 2015 from the UK Clinical Research Practice Datalink. Musculoskeletal symptoms identified by medcodes were concatenated into primary groups, each made up of several sub-groups. Baseline demographics for gender, age, body mass index (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 using: 1) primary musculoskeletal symptom groups, 2) musculoskeletal symptom sub-groups and 3) demographic variables. Predictive ability of the networks using the receiver operating characteristic curve (AUC) 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 sub-groups, was 76% accurate in predicting the development of PsA in a test set and had an AUC of 0.73 (95% confidence interval (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.

INTRODUCTION

Psoriasis is a common skin disorder which 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 psoriatic arthritis (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 osteoarthritis 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 nonspecific 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 noninflammatory rheumatic conditions, characterising the relationships between symptoms are made difficult by their highly-dependent nature, which violates the (conditional) independence assumption of most models. Approaches which 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 represent a set of variables and their conditional dependencies, in addition to being able to handle uncertainty. As such, in this paper we investigate their use in order to generate models which can be used to identify and characterise the relationships between musculoskeletal symptoms and the development PsA in people with psoriasis.

METHODS

Data source

This study used data from the Clinical Practice Research Datalink (CPRD), a UK electronic healthcare database, which contains anonymised longitudinal medical records for more than 20 million people.^(8,9) Patients 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 to 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 one 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 one 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 disease-modifying anti-rheumatic drug (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, body mass index (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 Organisation BMI categories: <25.0, 25.0-29.9, 30.0-34.9, \geq 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 three years before their psoriasis index date. Smoking status was classified as smoker, ex-smoker or non-smoker. Alcohol consumption status was categorised as heavy drinker (>6 units/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 (e.g. phototherapy, PUVA, or systemic therapies) or evidence of a referral to a dermatologist and 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 (NJM and WT) for suitability of inclusion. Redundant medcodes (medcodes which were unlikely to represent a musculoskeletal symptom associated with the development of PsA), along with those relating to other noninflammatory 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 sub-groups which were generally more site-specific (45 sub-groups in total). The medcode compositions of these six primary groups and 45 sub-groups are shown in Supplementary Table 1. Although some patients had multiple, identical symptom medcodes recorded within the study

period, binary features were created for all groups. While BN 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 characterise the relationships between baseline demographic variables, musculoskeletal symptoms and PsA. BNs are based on a graphical formalism called a Directed Acyclic Graph (DAG). 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 sub-groups and a subsequent BN network was constructed. For each BN, we utilised 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 data set, 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 (EM) to learn the BN structure from missing data.⁽¹⁷⁾ The EM 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 to 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 1 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 (AUC) were calculated using the test data set.

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 an 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 material.

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 summarised in *Error! Reference source not found.* The mean age at psoriasis index was 48 years (SD 18) and the mean duration of psoriasis was 5.71 (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 sub-groups (*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 2*). Over one fifth of the psoriasis patients who went on to develop PsA visited their GP with musculoskeletal-related symptoms during the five years prior to their diagnosis of PsA. This proportion gradually increased and reached over 57% in the six 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-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 to 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 six primary musculoskeletal symptom groups (arthritis and swelling), and ten of the 45 musculoskeletal symptom sub-groups (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 AUC was 0.73 (95% confidence interval (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 sub-groups. This BN was 76% accurate in predicting the development of PsA. The graphical representation of this BN structure is displayed in

Figure 1 and shows the widespread probabilistic associations between the variables included in the modelling. The graphical representations of the BN structures learnt using 1) the six primary musculoskeletal symptom groups and 2) only the demographic variables, can be found in the supplementary material (Supplementary Figure 1 and Supplementary Figure 2, 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 the valuable opportunity for the early detection of PsA. Whilst 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 paper 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 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 the 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 which are predecessors of PsA. Furthermore, including these musculoskeletal symptoms in our modelling improved the predictive ability of the BN when compare to 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 when it 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 nonspecific 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 which 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 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 as 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 EHR 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 rheumatologist diagnosis of PsA from 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 they would have referred 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 minimised 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 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 EM algorithm alongside clinical domain knowledge and published literature.⁽³²⁾ Last, 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 which 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 pre-clinical 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 characterising the relationships between musculoskeletal symptoms and the development 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 American College of Rheumatology 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/>

Competing interests

All authors have completed the ICMJE 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.

Contributors: NM contributed to the conception of the study. All authors contributed to the design of the work. Data acquisition and analysis was carried out by AG and TS. 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.

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.

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

Transparency declaration

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.

Data Availability

The data underlying this manuscript were provided by the Clinical Practice Research Datalink (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 listed used to identify musculoskeletal symptoms will be shared on reasonable request to the corresponding author.

REFERENCES

1. Menter A, Korman NJ, Elmets CA, Feldman SR, Gelfand JM, Gordon KB, 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(4):643-59.
2. Ogdie A, Weiss P. The Epidemiology of Psoriatic Arthritis. *Rheum Dis Clin North Am*. 2015;41(4):545-68.
3. Rosen CF, Mussani F, Chandran V, Eder L, Thavaneswaran A, Gladman DD. Patients with psoriatic arthritis have worse quality of life than those with psoriasis alone. *Rheumatology (Oxford)*. 2012;51(3):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. *Journal of Autoimmunity*. 2017;76:21-37.
5. Charlton RA, Tillett W, Nightingale AL, Snowball JM, Green ACA, Smith C, et al. 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, Vis M, Tchetverikov I, Nijsten TEC, et al. Prevalence of Psoriatic Arthritis in Primary Care Patients With Psoriasis. *Arthritis & Rheumatology*. 2016;68(4):924-31.
7. Eder L, Polachek A, Rosen CF, Chandran V, Cook R, Gladman DD. The Development of Psoriatic Arthritis in Patients With Psoriasis Is Preceded by a Period of Nonspecific Musculoskeletal Symptoms: A Prospective Cohort Study. *Arthritis & Rheumatology*. 2017;69(3):622-9.
8. Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, van Staa T, et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol*. 2015;44(3):827-36.
9. CPRD. Clinical Practice Research Datalink [Available from: www.cprd.com].
10. Chisholm J. The Read clinical classification. *British Medical Journal*. 1990;300(6732):1092-.
11. Green A, Shaddick G, Charlton R, Snowball J, Nightingale A, Smith C, et al. Modifiable risk factors and the development of psoriatic arthritis in people with psoriasis. *British Journal of Dermatology*. 2020;182(3):714-20.
12. Seminara NM, Abuabara K, Shin DB, Langan SM, Kimmel SE, Margolis D, et al. Validity of The Health Improvement Network (THIN) for the study of psoriasis. *Br J Dermatol*. 2011;164(3):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(7):443-51.
14. Ogdie A, Alehashemi S, Love TJ, Jiang Y, Haynes K, Hennessy S, et al. Validity of psoriatic arthritis and capture of disease modifying antirheumatic drugs in the health improvement network. *Pharmacoepidemiol Drug Saf*. 2014;23(9):918-22.
15. Jammeh EA, Carroll C, B, Pearson S, W, Escudero J, Anastasiou A, Zhao P, et al. Machine-learning based identification of undiagnosed dementia in primary care: a feasibility study. *BJGP Open*. 2018;2(2):bjgpopen18X101589.
16. Koller D FN, Bach F. Probabilistic graphical models: principles and techniques - Adaptive Computation and Machine Learning: Cambridge, MA: The MIT press; 2009.

17. Friedman N. Learning Belief Networks in the Presence of Missing Values and Hidden Variables. *Proceedings of the 14th International Conference on Machine Learning*. 1997:125–33.
18. Lauritzen SL. The EM algorithm for graphical association models with missing data. *Computational Statistics & Data Analysis*. 1995;19(2):191-201.
19. Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*. 1977;39(1):1-22.
20. Sugiyama K, Tagawa S, Toda M. Methods for Visual Understanding of Hierarchical System Structures. *IEEE Transactions on Systems, Man, and Cybernetics*. 1981;11(2):109-25.
21. John Ellson ERG, Eleftherios Koutsofios, Stephen C. North, and Gordon Woodhull. Graphviz and Dynagraph – Static and Dynamic Graph Drawing Tools [Available from: <https://www.graphviz.org/Documentation/EGKNW03.pdf>].
22. Farooq K, Hussain A, Leslie S, Eckl C, MacRae C, Slack W, editors. *An Ontology Driven and Bayesian Network Based Cardiovascular Decision Support Framework* 2012; Berlin, Heidelberg: Springer Berlin Heidelberg.
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 Physics Letters B*. 2017;31(19-21):1740055.
24. Seixas FL, Zadrozny, B., Laks, J., Conci, A., & Muchaluat Saade, D. C. A Bayesian network decision model for supporting the diagnosis of dementia, Alzheimer’s disease and mild cognitive impairment. *Computers in biology and medicine*. 2014;51:140-58.
25. Sambo F, Facchinetti A, Hakaste L, Kravic J, Di Camillo B, Fico G, et al., editors. *A Bayesian Network for Probabilistic Reasoning and Imputation of Missing Risk Factors in Type 2 Diabetes* 2015; Cham: Springer International Publishing.
26. Loghmanpour NA, Kanwar, M. K., Druzdzel, M. J., Benza, R. L., Murali, S., & Antaki, J. F. A new Bayesian network-based risk stratification model for prediction of short-term and long-term LVAD mortality. *ASAIO journal (American Society for Artificial Internal Organs : 1992)*. 2015;61(3):313–23.
27. Fuster-Parra P, Tauler, P., Bennasar-Veny, M., Ligeza, A., López-González, A. A., & Aguiló, A. Bayesian network modeling: A case study of an epidemiologic system analysis of cardiovascular risk. *Computer methods and programs in biomedicine*. 2016;126:128-42.
28. Gudu T, Etcheto A, de Wit M, Heiberg T, Maccarone M, Balanescu A, et al. Fatigue in psoriatic arthritis - a cross-sectional study of 246 patients from 13 countries. *Joint Bone Spine*. 2016;83(4):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 & Research*. 2015;67(2):264-72.
30. Orbai A-M, de Wit M, Mease P, Shea JA, Gossec L, Leung YY, et al. International patient and physician consensus on a psoriatic arthritis core outcome set for clinical trials. *Annals of the Rheumatic Diseases*. 2017;76(4):673.
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(1):40-5.

32. Lin J-H, Haug PJ. Exploiting missing clinical data in Bayesian network modeling for predicting medical problems. *Journal of Biomedical Informatics*. 2008;41(1):1-14.

TABLES AND FIGURES

Figure 1 Graphical model of the Bayesian network built using CPRD 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, as such, the order/importance of symptoms cannot be inferred from it.

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

| | All (n = 90189) | Psoriasis-only (n = 88780) | PsA (n = 1409) |
|---|--------------------|-------------------------------|-------------------|
| Sex*, no. (%) male | 43599 (48) | 42 854 (48) | 745 (53) |
| Age*, mean (SD), years | 48.3 (18.2) | 48.5 (18.2) | 44.7 (13.9) |
| Psoriasis duration**, mean (SD), years | 5.7 (4.1) | 5.8 (4.1) | 3.6 (3.4) |
| Psoriasis severity*, no. (%) severe | 11491 (13) | 11 083 (12) | 408 (29) |
| Musculoskeletal symptom†, no. (%) | 4571 (5) | 4404 (5) | 167 (12) |
| Arthritis | 109 (0) | 92 (0) | 17 (1) |
| Bursitis, Enthesistis 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***, mean (SD) | 4.5 (3.5) | 4.5 (3.5) | 3.5 (3.1) |
| BMI category‡, no. (%) (kg m 2) | | | |
| <25 | 15519 (17) | 15384 (17) | 135 (10) |
| 25.0-29.9 | 15841 (18) | 15603 (18) | 238 (17) |
| 30.0-34.9 | 8980 (10) | 8818 (10) | 162 (11) |
| 35>= | 5854 (6) | 5728 (6) | 126 (9) |
| Missing | 43995 (49) | 43247 (49) | 748 (53) |
| Smoking status‡, no. (%) | | | |
| Non-smoker | 38452 (43) | 37808 (43) | 644 (46) |
| Ex-smoker | 25039 (28) | 24646 (28) | 393 (28) |
| Current smoker | 25133 (28) | 24768 (28) | 365 (26) |
| Missing | 1564 (2) | 1558 (2) | 7 (0) |
| Alcohol status‡, no. (%) | | | |
| Non-drinker | 9883 (11) | 9745 (11) | 138 (10) |
| Ex-drinker | 4430 (5) | 4370 (5) | 60 (4) |
| Current-drinker | 63997 (71) | 62924 (71) | 1073 (76) |
| Missing | 11879 (13) | 11741 (13) | 138 (10) |

* on psoriasis index date, ** 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, *** calculated as the period between psoriasis diagnosis to symptom report date[†] ≥1 musculoskeletal symptom on psoriasis index date, † closest to and within 3 years prior to psoriasis index date. PsA, Psoriatic arthritis; BMI, Body Mass Index.

Table 2 Grouping of medcodes into musculoskeletal symptom groups

| Primary musculoskeletal symptom groups (n, Medcodes) | Musculoskeletal symptom sub-groups (n, Medcodes) |
|---|---|
| ARTHRITIS (35) | Ankle arthritis (1), Elbow arthritis (1), Foot arthritis (1), Hand arthritis (5), Hip arthritis (2), Knee arthritis (1), Osteoarthritis (7), Other arthritis (16), Wrist arthritis (1) |
| BURSITIS, ENTHESITIS OR TENDONITIS (73) | Bursitis (15), Enthesitis (35), Tendonitis (23) |
| FATIGUE (23) | Fatigue (22), Malaise (21) |
| PAIN (135) | Ankle pain (6), Arm pain (4), Arthralgia (13), Back pain (28), Chest pain (1), Elbow pain (5), Fibromyalgia (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), Sacroilitis (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 welling (2), Knee swelling (6), Shoulder swelling (3), Unspecified swelling (17) |

Table 3 Incidence of musculoskeletal symptoms

| | Cases | Person years | Incidence rate per 1000 person years (95% CI) | |
|---|-------|--------------|---|-------------------|
| All | | | | |
| Psoriasis | 53711 | 243230.691 | 220.82 | (218.96 - 222.69) |
| PsA | 1092 | 2104.038 | 519.00 | (488.22 - 549.79) |
| ARTHRITIS | | | | |
| Psoriasis | 1828 | 501928 | 3.64 | (3.47 - 3.81) |
| PsA | 188 | 4797 | 39.19 | (33.59 - 44.79) |
| BURSITIS, ENTHESITIS OR TENDONITIS | | | | |
| Psoriasis | 10119 | 465585 | 21.73 | (21.31 - 22.16) |
| PsA | 204 | 4567 | 44.67 | (38.54 - 50.8) |
| FATIGUE | | | | |
| Psoriasis | 11420 | 459804 | 24.84 | (24.38 - 25.29) |
| PsA | 127 | 4566 | 27.81 | (22.98 - 32.65) |
| PAIN | | | | |
| Psoriasis | 45374 | 290260 | 156.32 | (154.88 - 157.76) |
| PsA | 847 | 2759 | 307.04 | (286.36 - 327.72) |
| STIFFNESS | | | | |
| Psoriasis | 800 | 506921 | 1.58 | (1.47 - 1.69) |
| PsA | 41 | 4961 | 8.26 | (5.73 - 10.79) |
| SWELLING | | | | |
| Psoriasis | 6603 | 482902 | 13.67 | (13.34 - 14.00) |
| PsA | 207 | 4624 | 44.77 | (38.67 - 50.86) |

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

| Bayesian network | Accuracy | Sensitivity | Specificity | AUC |
|------------------|--------------------|--------------------|--------------------|--------------------|
| 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 sub-groups 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)

KEY MESSAGES

What is already known about this subject?

- Limited information of on the preclinical phase of PsA exists but there is evidence that a complex of non-specific musculoskeletal symptoms may precede the development of PsA in people with psoriasis.

What does this study add?

- Using a BN approach, we have shown that there are several musculoskeletal symptoms which are predecessors of PsA, including non-specific arthritis, fatigue, specific types of pain, and swelling.

How might this impact on clinical practice or future developments?

- 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.