BMJ Open

Early identification of persistent somatic symptoms in primary care: data-driven and theory-driven predictive modelling based on electronic medical records of Dutch general practices

Willeke M Kitselaar ,[1,2] Frederike L Büchner ,[1] Rosalie van der Vaart ,[2] Stephen P Sutch ,[1,3] Frank C Bennis ,[4,5] Andrea WM Evers ,[2] Mattijs E Numans [1]

To cite: Kitselaar WM, Büchner FL, van der Vaart R, *et al*. Early identification of persistent somatic symptoms in primary care: data-driven and theory-driven predictive modelling based on electronic medical records of Dutch general practices. *BMJ Open* 2023;**13**:e066183. doi:10.1136/bmjopen-2022-066183

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2022-066183).

Received 29 June 2022
Accepted 31 March 2023

For numbered affiliations see end of article.

Correspondence to
Willeke M Kitselaar;
w.m.kitselaar@vu.nl

ABSTRACT

Objective The present study aimed to early identify patients with persistent somatic symptoms (PSS) in primary care by exploring routine care data-based approaches.

Design/setting A cohort study based on routine primary care data from 76 general practices in the Netherlands was executed for predictive modelling.

Participants Inclusion of 94 440 adult patients was based on: at least 7-year general practice enrolment, having more than one symptom/disease registration and >10 consultations.

Methods Cases were selected based on the first PSS registration in 2017–2018. Candidate predictors were selected 2–5 years prior to PSS and categorised into data-driven approaches: symptoms/diseases, medications, referrals, sequential patterns and changing lab results; and theory-driven approaches: constructed factors based on literature and terminology in free text. Of these, 12 candidate predictor categories were formed and used to develop prediction models by cross-validated least absolute shrinkage and selection operator regression on 80% of the dataset. Derived models were internally validated on the remaining 20% of the dataset.

Results All models had comparable predictive values (area under the receiver operating characteristic curves=0.70 to 0.72). Predictors are related to genital complaints, specific symptoms (eg, digestive, fatigue and mood), healthcare utilisation, and number of complaints. Most fruitful predictor categories are literature-based and medications. Predictors often had overlapping constructs, such as digestive symptoms (symptom/disease codes) and drugs for anti-constipation (medication codes), indicating that registration is inconsistent between general practitioners (GPs).

Conclusions The findings indicate low to moderate diagnostic accuracy for early identification of PSS based on routine primary care data. Nonetheless, simple clinical decision rules based on structured symptom/disease or medication codes could possibly be an efficient way to support GPs in identifying patients at risk of PSS. A full data-based prediction currently appears to be hampered by inconsistent and missing registrations. Future research on predictive modelling of PSS using routine care data should focus on data enrichment or free-text mining to overcome inconsistent registrations and improve predictive accuracy.

STRENGTHS AND LIMITATIONS OF THIS STUDY

⇒ This is the first cohort study to apply predictive modelling 2 years prior to persistent somatic symptoms onset, based on a large sample size (n=94 440) and at least 7 years of temporal data.
⇒ This study used a wide range of predictors with high clinical relevance and generalisability to general practice.
⇒ Different data-driven and theory-driven approaches for identifying candidate predictors were employed and provide insight into the utility of different approaches.
⇒ The predictors' generalisability to the general population and interpretation should be done with caution since predictor registration depends on consulting and registration behaviour.

INTRODUCTION

In the general population, up to 10% of adults experience persistent somatic symptoms (PSS) that cannot be fully attributed to established biomedical pathological mechanisms.[1–4] PSS are present in both patients with well-established diseases, such as cancer[5] and cardiovascular disease,[6] and in patients with symptoms without well-established biomedical pathology.[1] PSS are not only burdensome to the patient,[7] but also greatly impact healthcare.[8] For instance, in general practice, up to 50% of consultations are related to symptoms which are not clearly relatable to biomedical pathology.[9] Most of these symptoms are self-limiting and do not need further

investigation or treatment. However, identifying patients at risk of developing persistent symptoms is generally challenging.[10]

Definitions of PSS are ever-changing. Historically, PSS classification was based on the exclusion of well-established physical conditions.[11] Recent developments lack such a distinction and focus on more positive definitions (including dysfunctional symptom perceptions).[12 13] Moreover, PSS may be defined under broad 'umbrella' terms or based on specific syndromes such as irritable bowel syndrome (IBS), fibromyalgia (FM) or chronic fatigue syndrome (CFS). Previous research debated the distinctness of specific syndromes.[14] However, nowadays most experts accept accumulating evidence that there are both overarching common factors as well as syndrome-specific aspects to PSS.[15 16] Similarly, differing terminology is used between healthcare professionals. For instance, in psychiatry the umbrella term 'somatic symptom disorder' may be used, whereas in general medicine the term 'functional somatic symptoms' is used.[13 17 18] Lastly, some physicians refrain from using terms beyond well-established biomedical disorders for somatic symptoms.[19 20] In this paper, we use the term PSS, since we aim to approach identifying the broad spectrum of patients with persistent symptoms without well-established pathophysiology, and since recent research indicates that this term is generally preferred over other umbrella terms.[21]

Ambiguity in definitions and terminology has contributed to hampered (early) identification and proactive clinical intervention of patients at risk of developing PSS.[22–24] For instance, research shows that patients with fibromyalgia are diagnosed around 6 years after symptom onset.[25] Consequently, PSS are related to inappropriate and relatively high healthcare utilisation and costs.[26–28] Especially in many Western countries, where general practitioners (GPs) serve as a gatekeeper for specialist healthcare.[29 30] To prevent unnecessary referrals and medicalisation, with potential risk of iatrogenic harm, and to enable the initiation of proactive interventions, early identification is necessary.[31 32] However, there are many barriers towards the identification of PSS in primary care.[10 19] For example, diagnosis may be difficult due to the predominance of the biomedical disease model, fear of missing malignancy or other life-threatening conditions, the GP's experience and knowledge relating to PSS and consultation constraints like overloaded surgery hours. Research from a European network of experts in the field stresses the need for a systemic change to overcome these challenges.[33] Furthermore, research shows that an integrative care approach (with attention to psychological, social, interpersonal and contextual factors, in addition to keeping track of any biomedical deterioration) is needed to improve care for PSS.[34 35]

Over the years, several screening tools for patients with PSS-related issues were developed for clinical use.[1 36–38] While diagnostic accuracy and validity have been demonstrated, the widespread use is not forthcoming. A survey of Dutch GPs showed that GPs are still in need of tools for PSS-related diagnostics.[20] Studies have shown that routine care data can be responsibly used for predictive modelling.[39 40] The development of prediction models based on routine primary care data may enable screening based on readily available clinical information and support GPs in their practice. Recent studies reveal the multi-applicability of routine care data since it can be used in several different ways. Approaches range from the more classic theory-driven approaches, simple data-driven approaches[41] and more complex temporal data-mining techniques.[39 40]

This paper represents the first attempt to develop a clinical decision rule for PSS onset based on routine primary care data. The study aims to predict what patients are at risk of developing PSS 2 years prior to onset and explores different candidate predictor selection approaches. While a theory-driven approach is well-established and has a long history in science, especially in cohort studies, the use of routine care data potentially provides an approach that is more generalisable to clinical practice. Moreover, since we cannot control variable collection, we are interested in how theory-driven variable selection performs compared with non-routinely collected studies. Therefore, the present study explores different theory and data-driven approaches of variable selection, and their combinations, to identify the best approach for the predictive modelling of PSS.

METHODS

Study design

A population-based retrospective cohort study was performed using data from 76 primary care practices affiliated with the extramural Leiden academic network (ELAN) of the Leiden University Medical Center (LUMC), the Netherlands. First, the onset date of PSS was determined according to the approach described below (see the Outcome section) within the period 1 January 2017 until 31 December 2018 (random 'onset' dates were selected for patients without PSS). Thereafter, candidate predictors were selected 2–7 years prior to the onset date (ie, for each patient 5 years of data was used to select candidate predictors). The ELAN data consists of several subsets, including demographic data (gender, year of birth), consultations (dates, coded symptomology and diagnoses according to the Dutch version of the WONCA International Classification of Primary Care (ICPC)[42]), prescribed medication (dates and coded WHO anatomical therapeutic chemical (ATC) classification[43]), laboratory test (dates and results) and correspondence data (dates and type of healthcare professionals (eg, profession/specialty of the other professional).[44] Part of the consultation registration is the ICPC-coded episode registration, where chronic disorders are registered. The episode data may be available up to the date of birth.

Study population

Patients aged 25–100 years from the ELAN data warehouse were used for this study. Participating practices were located in the greater Leiden and The Hague area. In general, all Dutch residents are enlisted and registered at a general practice in their neighbourhood. Primary care is included in the mandatory Dutch insurance and free of additional charge for insured citizens. The ELAN data warehouse consists of pseudonymised routine healthcare data extracted from the electronic medical records (EMRs).[45] Inclusion criteria were: registered at the general practice for at least 7 years, having at least 10 contacts and 1 ICPC code. These criteria were used to ensure availability of enough registrations per patient to enable candidate predictor construction. Furthermore, due to higher likelihood of registration errors, patients who were over 100 years of age on 31 December 2018 were excluded from the study. Because we were interested in PSS onset prediction, patients who were registered with PSS before 1 January 2017 were excluded from the analysis.

Outcome

The definition of PSS is based on an earlier analysis by our research group, for which the same ELAN database was used.[32] Three approaches towards PSS identification were applied. Patients were identified as having PSS based on either having (1) ICPC codes for PSS syndromes (A04.01: chronic fatigue syndrome, D93: irritable bowel syndrome and L18.01: fibromyalgia); (2) PSS-umbrella terms, PSS-syndrome or PSS-complaint description in the episode description and/or (3) a score of ≥20 on the somatisation subscale of the four-dimensional symptom questionnaire (4DSQ), registered in the lab results. For a more detailed description of the selection criteria, see Kitselaar *et al*.[32]

Candidate predictors

Different datasets were constructed with specific theory and data-driven candidate predictors of PSS in the ELAN data. Below a brief description of the predictor categories related to each dataset-based model will be given, see figure 1 for an overview of the data extraction steps and online supplemental table S1 for a detailed overview of candidate predictors. Two distinct theory-driven datasets were operationalised: (1) literature-based risk factors of PSS (see Kitselaar *et al*[35] for more detail) and (2) frequencies of specific PSS-related terms and words in the free text with limited structured registration options (see online supplemental table S1). Data-driven datasets were divided into non-temporal and temporal data-driven datasets. The non-temporal datasets consist of dichotomised medical coding data (symptom/disease codes, medication codes and referrals). The coded symptom/disease dataset was based on ICPC codes categorised into WONCA chapters and code categories.[46] The coded medication dataset was based on ATC codes reduced to third level (to therapeutic/pharmacological subgroup[47]). The referral dataset was based on correspondences GPs have with other healthcare professionals.

The temporal approach consists of contextualised lab results and sequential patterns in medical coding data. Due to the high number of different lab results and

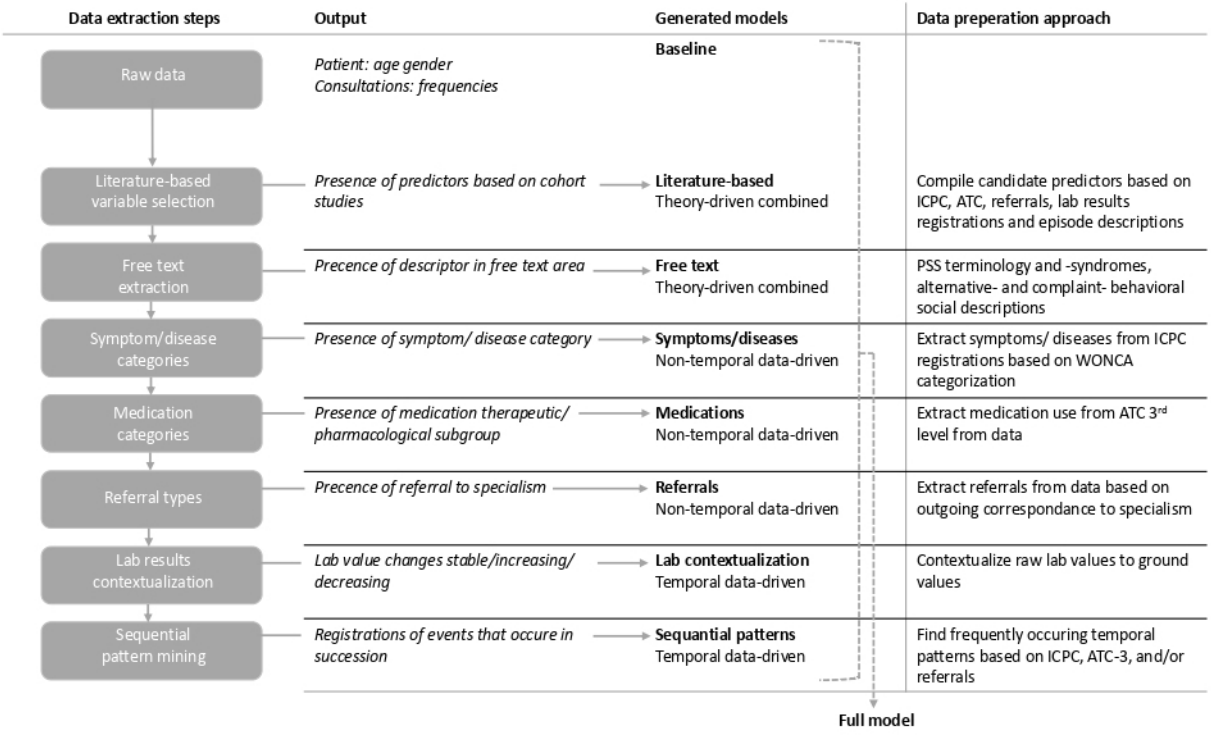

Figure 1 Diagram showing the data extraction steps for each constructed model. ATC, Anatomical therapeutic Chemical classification; ICPC, International Classification of Primary Care; PSS, persistent somatic symptoms.

inconsistent availability, using reference values for this study was not feasible. Contextualisation of lab results provides a solution to enable interpretability of lab results for individual patients. In relative grounding, a lab value is comparted to its previous value to deter whether values are decreasing, increasing or have remained stable.[39] To avoid relatively small fluctuations in lab values as decreases or increases, variables were scaled and a minimum of 5% difference between values was required to count as a change. After relative grounding the number of stable, decreased and increased values per lab measure were used as candidate predictors.

Sequential pattern identification of medical coding data was detected using the Sequential PAttern Discovery using Equivalence classes (SPADE) algorithm.[48] The SPADE algorithm is an efficient way to find statistically significant patterns in temporal data. To identify patterns with the SPADE algorithm, sequences of registrations (ICPC, ATC and referrals) are ordered by date and subsequent registrations are associated to each object in which it occurs.[48] Thus, when a patient has multiple registrations on one day these will be separated and combined with possible subsequent registrations (eg, patient X has the following registrations on date Y: fatigue, abdominal pain, anti-constipation drug and date Z: physiotherapy, this will result in three patterns for patient X: (1) fatigue→physiotherapy; (2) abdominal pain→physiotherapy and (3) anti-constipation drug→physiotherapy). We selected frequent patterns as candidate predictors based on having at least 1% difference between patients with PSS and patients without PSS in the support value (ie, prevalence of the pattern in de dataset). Please see Zaki[48] for a more detailed description of the SPADE algorithm.

Predictive modelling

For predictive modelling, a machine learning approach by means of least absolute shrinkage and selection operator (LASSO) logistic regression was used. Relating to our dataset and aim, LASSO logistic regression has several advantages over other methods. LASSO is especially suitable for unbalanced datasets, in which the outcome classification groups differ greatly in size. Moreover, LASSO avoids overfitting in in case of a great number of candidate predictors[49] and when multicollinearity is expected.[50] Regression was chosen because of its general comprehensibility and because previous studies in EMR data have shown this generally preforms all popular methods.[39 51]

The combined dataset was stratified into a training set (80%) and test set (20%). For training, a fivefold cross-validation, with hyperparameter tuning, was performed on the training set. For each unique model (ie, literature-review, free text, coded symptom/diseases, coded medications, referrals, contextualisation of lab results and sequential patterns) and all combined models (ie, theory-driven, data-driven non-temporal, data-driven temporal and full model), near zero-variance candidate predictors were removed (see online supplemental table S2 for total number of candidate predictors in the model and data sources). To evaluate the predictive value of each model, a sensitivity analysis was performed. This included prevalence independent measures (ie, sensitivity and specificity) and prevalence dependent measures (ie, positive predictive value (PPV) and negative predictive value (NPV)). Notably, PPV and NPV should be interpreted with caution because they are generally low when prevalence is low and their value is debatable when the prevalence in the study is not similar to general population prevalence (for a more detailed description, see [52 53]). Finally, the area under the receiver operating characteristic curve (AUC) was calculated. All data was prepared and analysed using R v4.0. For the final modelling, the caret-package was used.

Final model evaluation

To evaluate the models obtained using from model training (using the training dataset) and ensure there was no overfitting of the models, the models were internally validated on the test dataset for their classification performance. Finally, predictors of the final full model were evaluated. Estimated coefficients of predictors included in the final model were presented as ORs. To verify the stability of the predictor estimates, frequencies of estimates receiving non-zero values were calculated across 1000 bootstrap samples.

Role of the funding source

The project was internally funded by the Leiden University and Leiden University Medical Centre interdisciplinary profile area 'Health Prevention and the Human Life Cycle'. No external funding supported this study.

Patient and public involvement

GPs affiliated with the LUMC health campus were consulted during the development phase of the research design. Meetings with GPs were directed at the formulation of the outcome and construction of candidate predictors. Primary focus was the meaning and application of ICPC codes, lab measures, likelihood of missing data and general workings of EMR. Also locations to find relevant resources were discussed, to increase the knowledge of the data and the best way to interpret registrations.

RESULTS

The total number of patients in the ELAN database we used for our research contained 306 859 patients, of which a total of 202 168 patients were excluded based on available data. A total of 10 249 patients were classified as having PSS before 1 January and therefore also excluded from the study. As a result, 94 440 patients were included in the final analysis (figure 2).

As shown in table 1, 0.9% (n=902) of patients in the ELAN cohort had new-onset PSS. Compared with the total cohort, patients with PSS are more likely to be women (69.0% vs 52.9%, p<0.001), are generally younger (52.6±14.4 vs. 57.2±15.4, p<0.001) and have higher

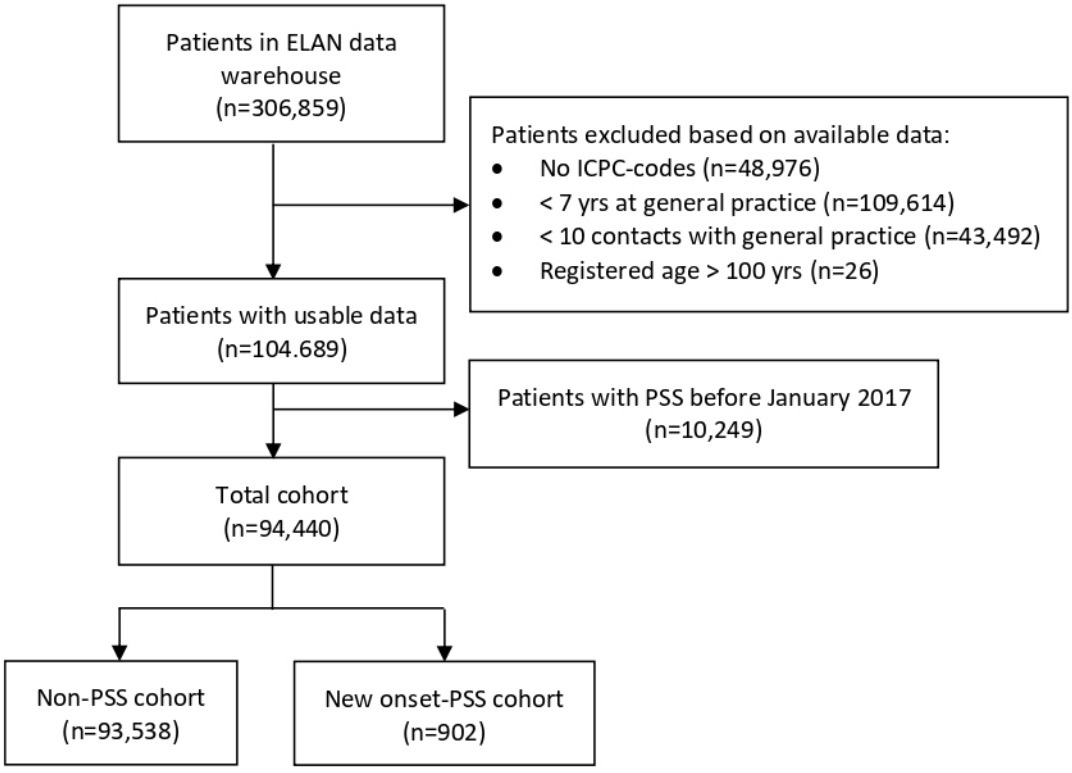

Figure 2 Flow chart of patient inclusion in the ELAN study cohort. ELAN, extramural Leiden academic network; ICPC, International Classification of Primary Care; PSS, persistent somatic symptoms.

consultation frequency (8.7±7.3 vs. 6.3±5.8, p<0.001). Moreover, patients with PSS are more likely to have a mental health disorder (60.3% vs 46.8%, p<0.001) while the likelihood of a physical disorder does not differ (64.6% vs 63.6%, p=0.87). The patients with new-onset PSS in the training and test sets differ on baseline variable women (68.3% vs 72.2%). Post-hoc evaluation revealed that patients with PSS in the training and test sets also differ regarding the prevalence of mental comorbidities (59.6% vs 63.3%, respectively) and physical comorbidities (65.1% vs 62.8%) (not depicted in table).

In table 2, the predictive value based on sensitivity, specificity and the AUCs of each unique and combined model is depicted. The AUCs of the validated models varied from 0.68 for the baseline model to 0.72 for the full model. From the separate models, all models preformed equally well, based on an approximate AUC 0.70. PPV is low (ranging from 1.5% to 1.7%) and NPV is high (ranging

Table 1 Patient characteristics

	Total cohort	PSS		
	Full dataset	Full dataset	Training	Test
n (%)	94 440 (100.00)	902 (0.9)	772 (0.9)	180 (0.9)
Female, n (%)	49 998 (52.9)	623 (69.0)	493 (68.3)	130 (72.2)
Age, mean (SD)	57.2 (15.4)	52.6 (14.4)	52.9 (14.5)	51.3 (13.7)
Consultations, mean (SD)	6.3 (5.8)	8.7 (7.3)	7.44 (6.3)	7.2 (5.5)
Urbanisation level, n (%)				
Urban area	45 567 (48.2)	404 (44.8)	326 (45.2)	78 (43.3)
Sub-urban area	43 296 (45.8)	448 (49.7)	358 (49.6)	90 (50.0)
Rural	2711 (2.9)	9 (1.0)	7 (1.0)	2 (1.1)
Disadvantage neighbourhood	67 215 (71.2)	622 (69.0)	494 (68.4)	128 (71.1)
Physical comorbidity, n (%)	60 019 (63.6)	583 (64.6)	470 (65.1)	113 (62.8)
Mental comorbidity, n (%)	44 292 (46.9)	544 (60.3)	430 (59.6)	114 (63.3)

PSS, persistent somatic symptoms.

Table 2 Prediction models based on LASSO logistic regression analysis

		Training	Test		
		AUC	Sensitivity	Specificity	AUC
	Baseline model*	0.66	0.73	0.54	0.68
Theory-driven	Literature-based†‡	0.70	0.61	0.68	0.71
	Free text†§	0.68	0.70	0.56	0.71
	Combined*	0.69	0.73	0.60	0.71
Non-temporal data-driven	Symptoms/diseases†¶	0.68	0.72	0.57	0.70
	Medications†**	0.69	0.76	0.58	0.70
	Referrals†††	0.66	0.71	0.55	0.69
	Combined†	0.70	0.57	0.72	0.71
Temporal data-driven	Lab contextualisation†‡‡	0.67	0.73	0.58	0.70
	Sequential patterns†§§	0.66	0.83	0.43	0.69
	Combined†	0.68	0.73	0.58	0.70
	Full model†¶¶	0.70	0.72	0.60	0.72

For a detailed description of the models, see online supplemental table S1.
*Gender, age and consultation frequency.
†It includes baseline model.
‡Variables selected based on literature search of risk factors in the general population.
§Word search through free journal text.
¶ICPC codes categorised according to the WONCA categorisation (dichotomised).
**ATC-3: therapeutic/pharmacological subgroup (dichotomised).
††Outgoing correspondence to medical specialists (dichotomised).
‡‡Relative grounded lab-results (stable, increase, decrease; dichotomised).
§§Order of ICPC, ATC and referrals over time, patterns identified with the SPADE algorithm (see online supplemental table S3).
¶¶All available candidate predictors combined; For a detailed description of the models, see online supplemental table S1.
ATC, anatomical therapeutic chemical; AUC, area under the receiver operating characteristic curve; ICPC, International Classification of Primary Care; LASSO, least absolute shrinkage and selection operator.

99.5% to 99.6%). Using the optimal cut-off selection (ie, highest number of cases selected accurately), the present model would, with 72.2% sensitivity, detect patients at-risk of PSS onset within 2 years (see table 2 for AUC's and sensitivity analyses, and online supplemental tables S1–S3 for more details on the model contents).

Final predictors were derived from the full model. From all candidate predictors used for the full model (n=545), 29 of the variables contributed to the prediction of PSS onset. Predictors stemmed from all predictor type categories, baseline (n=2), literature review (n=8), ATC (n=8), ICPC (n=3), free text (n=2), referrals (n=1), lab contextualisation (n=3) and sequential patterns (n=1). From the baseline predictors, age decreased the likelihood of PSS onset (OR=0.82) and female gender increased (OR=1.13) the likelihood of PSS onset. Baseline variable consultation frequency was not a relevant predictor in the full model, but it was an important predictor in all other models, except for the theory-driven combined model. Some other highly stable predictors using PSS-related complaint description in the free text (OR=1.12) are: having stable lymphocyte counts based on lab tests (OR=84.2); using PSS-related terminology in free text (OR=83.6%); the number of referrals for imaging (OR=1.10); number of medications (OR=1.12) and having a neurological disorder (OR=1.10) (see table 3 for the complete list of predictors and ORs). Frequencies of estimates having non-zero values across 1000 bootstrap samples indicate the level of interchangeability of predictors for other predictors (high percentage indicating higher importance of the predictor for predicting PSS onset).

Several of the predictors may have overlapping aetiology or overlapping variable constructs but differ in their data source. This is for instance seen in: (1) female genital symptoms (ICPC), painful intercourse (literature review), both contain ICPC code X04; (2) 'headache' (literature review) and neurological disorders (ICPC), both containing ICPC codes N89 and N90; (3) digestive symptoms (ICPC) and drugs for anti-constipation (ATC) and (4) 'fatigue' (ICPC) and 'complaint description' (free-text descriptors, which contains the term fatigue).

DISCUSSION

This study provides a comprehensive overview of the effectiveness of different approaches towards predicting PSS based on routine primary care data 2 years prior to index date. Model performance based on specific predictor generation approaches does not differ greatly. Therefore, the use of the simplest approach may be most desirable. Based on the full model (including all candidate

Table 3 Predictors of PSS obtained from full model LASSO logistic regression analysis

Predictors	Total cohort % or mean (SD)	PSS cohort % or mean (SD)	OR	%*
Baseline				
Age	57.2 (15.4)	52.6 (14.4)	0.82	99.5
Female gender	52.9	69.0	1.13	78.1
Literature based (theory-driven)				
Painful intercourse (female)†	1.1	3.1	1.17	60.8
Medications‡	2.0 (1.4)	2.5 (1.6)	1.12	94.7
Number of imaging referrals§	0.09 (0.09)	0.1 (0.1)	1.10	96.1
Fatigue¶	20.5	31.2	1.04	47.5
Mood disorder**	14.6	23.6	1.03	47.7
Number of pain sites††	0.3 (0.6)	0.5 (0.7)	1.02	63.7
Headache§§	19.8	32.6	1.02	44.8
Number of ICPC codes‡‡	2.6 (1.5)	3.3 (1.7)	1.004	13.5
Free text (theory-driven)				
Complaint description¶¶	0.7 (1.0)	1.3 (1.6)	1.12	99.3
PSS terminology***	0.06 (0.15)	0.11 (0.21)	1.04	83.6
Symptom/disease codes (non-temporal data-driven)				
Neurological disorder†††	18.1	27.3	1.11	77.9
Digestive symptoms‡‡‡	50.4	65.5	1.07	66.7
Female genital symptoms§§§	28.8	46.6	1.07	53.0
Female genital infection¶¶¶	8.3	15.9	1.04	48.9
Medication codes (non-temporal data-driven)				
Capillary stabilisers****	0.1	0.7	1.47	57.6
Selective CA+ blockers††††	10.6	6.3	0.93	58.0
Topical contraceptives‡‡‡‡	5.5	10.5	1.06	58.8
Lipid modifier§§§§	21.4	15.6	0.95	54.2
Nasal spray, topical¶¶¶¶	40.1	51.7	1.02	51.1
Anti-constipation drug*****	28.4	40.1	1.02	52.1
Eyedrops, topical†††††	16.2	22.3	1.01	47.3
Anti-thrombotic agents‡‡‡‡‡	20.8	16.0	0.999	41.0
Referrals (non-temporal data-driven)				
Physiotherapy§§§§§	30.2	39.5	1.01	43.6
Lab contextualisation (temporal data-driven)				
Lymphocytes, stable	0.3 (0.5)	0.4 (0.5)	1.06	84.2
Thyroid, stable	0.5 (1.1)	0.8 (1.4)	1.04	70.3
Systolic blood pressure, stable	1.8 (3.2)	1.5 (2.8)	0.999	39.0
Sequential patterns (temporal data-driven)				
Referral to Rontgen	3.1	7.1	1.10	57.6

Continued

Table 3 Continued

Predictors	Total cohort % or mean (SD)	PSS cohort % or mean (SD)	OR	%*

*Frequency of estimates having non-zero values across 1000 bootstrap samples
†ICPC codes: X04, P08.02.
‡Frequency based on full ATC codes.
§Rontgen or echography.
¶ICPC code: A04.
**ICPC codes: P03, P73, P73.02, P76 and ATC codes: N06A, N05AN, D11A×04.
††Number of pain-related ICPC codes.
‡‡ICPC codes: N01, N02, N89, N90, R09.
§§All unique ICPC codes.
¶¶Fatigue, dizziness, back pain (see online supplemental table S1) for full list).
***For example, somatisation or a-specific symptoms (see online supplemental table S1) for full list).
†††ICPC: N86-99.
‡‡‡ICPC codes: D01-29.
§§§ICPC codes: X01-29.
¶¶¶ICPC codes: X70-74 and X90-92.
****ATC4-codes: C05C.
††††ATC4 codes: C08C.
‡‡‡‡ATC4 codes: G02B.
§§§§ATC4 codes: C10A.
¶¶¶¶ATC4 codes: R01A.
*****ATC4 codes: A06A.
†††††ATC4 codes: S01X.
‡‡‡‡‡ATC4 codes: B01A.
§§§§§Correspondence with physiotherapy.
ATC, Anatomical Therapeutic Chemical classification; ICPC, International Classification of Primary Care; LASSO, least absolute shrinkage and selection operator; PSS, persistent somatic symptoms.

predictors), predictors associated with PSS onset stem from all predictor categories, although theory-driven and medication types (ATC) predictors were most prevalent. In line with previous literature, important predictors are related to being female (including, painful intercourse, genital infections/symptoms and contraceptives), specific symptoms (eg, digestive issues, fatigue, mood disorders and headache), healthcare utilisation (eg, number of medications or imaging, referrals or physiotherapy) and number of complaints (eg, number of pain sites or ICPC codes). Consistent with knowledge that PSS is unrelated to established biomedical pathology, results show that stable lab results (especially lymphocytes and thyroid) are important indicators of PSS. Notably, constructs of some predictors contain overlapping variables (such as: 'neurological disorder' and 'headache'; and 'fatigue' and 'complaint description'). This indicates that ambiguous registration may result in scattered predictors, which may have contributed to the limited predictive accuracy of the models.

Several strengths and limitations apply to this study. A major strength is the population-based cohort, with high ecological validity, with a large sample size and at least 7 years of data. Second, inclusion in our PSS cohort is based on a previously published approach which has enabled us to select patients beyond the poorly reported ICPC codes for the syndromes,[32] and not limited to commonly investigated IBS, FM and CFS.[54] To our knowledge, we included a wider range of predictors than previous studies, and these are clinically relevant and generalisable to general practice. Moreover, the models were compared based on predictor categories which provides important evidence for more efficient future analyses. Lastly, we have used sophisticated machine learning techniques (temporal pattern mining and relative grounding) and analysis (LASSO regression). This allowed for optimal use of temporal data and enabled us to use all available candidate predictors in one final model. Finally, although the machine learning techniques did not improve the performance of the full model, some novel predictors were identified (ie, stable lab results: lymphocytes and thyroid). On the other hand, the use of routine care data may also limit the generalisability of the predictors to the general population since registration depends on the decision of patients to contact the physician and on the decision of physician/staff what to register. Furthermore, interpretation of predictors should be done with caution since the present analysis is directed at finding the optimal model performance, rather than explaining the outcome. For example, registration of social and psychological predictors may frequently be missing, since medical priorities might be estimated as the more important issues to code and register.[32 41 52] Finally, the selection of patients with PSS was based on previous research on the same dataset.[32] This approach enabled conservative selection of patients with PSS, but may have missed some cases.[53 55] The aim was to enable data-driven selection and not rely on GP diagnosis, since research indicates that PSS are often missed by physicians.[56] Data-driven selection would enhance re-usability of routine care data.

To our knowledge, this is the first cohort study to predict PSS 2 years prior to onset. However, previous predictive EMR studies on PSS or PSS-subgroups show better model performance. This may be due to the 2-year prediction gap, which was not applied in previous studies or because of their use of questionnaires or physician-dependent diagnoses.[57–59] A recent study based on the ELAN data warehouse with a non-biomedical outcome showed similar predictive value,[41] which could mean that routine primary care data have limited capacity for non-biomedical outcome measures. However, this study also did not apply a 2-year prediction gap. Prediction models based on other types of large cohort studies have primarily focused on PSS subtypes.[54 59] Monden *et al*[54] reported notably higher ORs, which may be related to less available confounding variables and/or to active data collection resulting in access to multidomain (ie, more complete social and psychological) data. This is in line with studies showing that GPs are less likely to report social and psychological factors[19 20 60] and a recent systematic review demonstrating the importance of using multidomain data.[35] Lastly, in contrast to a body of evidence,[59 61 62] our LASSO regression of the full model did not indicate that consultation frequency predicts PSS. Since consultation frequency was predictive in most submodels, findings imply that factors latent to consultation (such as number of imaging referrals or number of ICPC-codes) may be more precise predictors of PSS onset than consultation frequency.

Our study shows how routine primary care data can be used as a source that supports early prediction of PSS. Although predictive accuracy (in particular shown by the low PPV) indicates that it cannot be used without additional screening, relatively simple ICPC/ATC-based models can assist in this process by facilitating an initial broad distinction between PSS and well-established biomedical problems. Predictive values of free text 'complaint description' and 'PSS terminology' indicate that clinical evaluation and registration of PSS-related psychological and social constructs is important for early identification of PSS. Thus, in combination with the simple ICPC/ATC-based models, available validated screening tools such as the 4DSQ and the somatic symptom disorder - B-criteria scale (SSD-12) might further facilitate early identification of PSS. Moreover, the overlapping constructs of several predictors, which do not correlate highly, indicate a difference in registration behaviour between GPs practices, which may have limited the predictive value of the data. Although sequential patterns and lab contextualisation did not enhance model performance, the former implies that other machine learning techniques (eg, text mining) should be further explored. Especially because of the relatively fair performance of the free text-based model, for which in the present study only limited free text is used.

Results provide clear directions for both clinical and EMR research. Clinical research should be directed at the feasibility of the ICPC/ATC-based models for clinical implementation in combination with additional screening with a validated screening tool (eg, 4DSQ or SSD-12). The screening tools would provide a proxy for the difficulty to systematically register PSS-related aspects captured in the free text. Future research should evaluate criterium validity of the present outcome by selecting the outcome (ie, PSS) using validated screening tools (eg, 4DSQ, SSD-12), and further evaluate if this could enhance accuracy of routine primary care data-based predictions. Furthermore, EMR research should further develop the theory-driven and data-driven approaches. The theory-driven approach could thus be improved by more elaborate candidate predictor construction, combining variables with similar constructs more thoroughly and patient-reported outcome measures. The data-driven approach could possibly be improved using data enrichment techniques or by developing models based on more advanced approaches for free-text analysis.

Author affiliations
[1]Health Campus The Hague/Department of Public Health and Primary Care, Leiden University Medical Center, The Hague, The Netherlands
[2]Health, Medical and Neuropsychology unit, Department of Psychology, Leiden University, Leiden, Netherlands
[3]HSR, Johns Hopkins University Bloomberg School of Public Health, Baltimore, Maryland, USA
[4]Computer Science, Vrije Universiteit Amsterdam, Amsterdam, Netherlands
[5]Netherlands Institute for Health Services Research, Utrecht, Netherlands

Twitter Stephen P Sutch @stevesutch

Acknowledgements The authors thank Dr Frank de Vos (Leiden University) for his advice and support regarding methodology and the verification of the stability of the predictors.

Contributors The study was primarily designed by WMK, RvdV, AWE and MEN. WMK conducted the study under the guidance of all other authors. WMK pre-processed the data. WMK analysed the data under the guidance of FCB, SPS and FLB. WMK drafted the manuscript. FLB, RvdV, FCB and SS reviewed and provided critical comments on all early-stage drafts of the manuscript. AWE and MEN reviewed and the provided critical comments on drafts of the full manuscript. All authors approved the submitted version. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted. MEN is the guarantor of this study.

Funding WMK's PhD project was internally funded by Leiden University (a profile area) and Leiden University Medical Center. No external funding supported this study.

Competing interests None declared.

Patient and public involvement Patients and/or the public were involved in the design, or conduct, or reporting, or dissemination plans of this research. Refer to the Methods section for further details.

Patient consent for publication Not applicable.

Ethics approval This study involves human participants. The ethics committee of Leiden University Medical Centre supplied a waiver of ethical approval (G19.045/SB/ib), as ethical approval was not necessary for this study, exempted this study. Since there was no active data collection for this study, it was not possible to ask participants personally for informed consent. Patients were given the option to opt-out of research participation in general by their general practice. All EMR data were transferred to a trusted third party that excluded patients who opted out of research participation. The rest of the patients were pseudonymised before transferring the data to us for analysis.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. No additional raw data are available. Processed data can be made available to researchers upon reasonable request.

ORCID iDs
Willeke M Kitselaar http://orcid.org/0000-0003-0914-0168
Frederike L Büchner http://orcid.org/0000-0001-8977-5344
Rosalie van der Vaart http://orcid.org/0000-0002-8787-4186
Stephen P Sutch http://orcid.org/0000-0003-3202-6718
Frank C Bennis http://orcid.org/0000-0002-6233-9101
Andrea WM Evers http://orcid.org/0000-0002-0090-5091
Mattijs E Numans http://orcid.org/0000-0002-0368-5426

REFERENCES

1 Kop WJ, Toussaint A, Mols F, *et al*. Somatic symptom disorder in the general population: associations with medical status and health care utilization using the SSD-12. *Gen Hosp Psychiatry* 2019;56:36–41.
2 Rief W, Burton C, Frostholm L, *et al*. Core outcome domains for clinical trials on somatic symptom disorder, bodily distress disorder, and functional somatic syndromes: European network on somatic symptom disorders recommendations. *Psychosom Med* 2017;79:1008–15.
3 Petersen MW, Schröder A, Jørgensen T, *et al*. Irritable bowel, chronic widespread pain, chronic fatigue and related syndromes are prevalent and highly overlapping in the general population: danfund. *Sci Rep* 2020;10:3273.
4 Katon W, Lin EHB, Kroenke K. The association of depression and anxiety with medical symptom burden in patients with chronic medical illness. *General Hospital Psychiatry* 2007;29:147–55.
5 Grassi L, Caruso R, Nanni MG. Somatization and somatic symptom presentation in cancer: A neglected area. *Int Rev Psychiatry* 2013;25:41–51.
6 Kohlmann S, Gierk B, Hümmelgen M, *et al*. Somatic symptoms in patients with coronary heart disease: prevalence, risk factors, and quality of life. *JAMA Intern Med* 2013;173:1469–71.
7 Choy E, Perrot S, Leon T, *et al*. A patient survey of the impact of fibromyalgia and the journey to diagnosis. *BMC Health Serv Res* 2010;10:102.
8 Burton C, Fink P, Henningsen P, *et al*. Functional somatic disorders: discussion paper for a new common classification for research and clinical use. *BMC Med* 2020;18:34.
9 Haller H, Cramer H, Lauche R, *et al*. Somatoform disorders and medically unexplained symptoms in primary care: A systematic review and meta-analysis of prevalence. *Dtsch Arztebl Int* 2015;112:279.
10 Murray AM, Toussaint A, Althaus A, *et al*. The challenge of diagnosing non-specific, functional, and somatoform disorders: A systematic review of barriers to diagnosis in primary care. *J Psychosom Res* 2016;80:1–10.
11 De Gucht V, Fischler B. Somatization: a critical review of conceptual and methodological issues. *Psychosomatics* 2002;43:1–9.
12 Rief W, Martin A. How to use the new DSM-5 somatic symptom disorder diagnosis in research and practice: A critical evaluation and A proposal for modifications. *Annu Rev Clin Psychol* 2014;10:339–67.
13 Löwe B, Mundt C, Herzog W, *et al*. Validity of current somatoform disorder diagnoses: perspectives for classification in DSM-V and ICD-11. *Psychopathology* 2008;41:4–9.
14 Chalder T, Willis C. "lumping'' and "splitting'' medically unexplained symptoms: is there a role for a transdiagnostic approach? *Journal of Mental Health* 2017;26:187–91.
15 Witthöft M, Fischer S, Jasper F, *et al*. Clarifying the latent structure and correlates of somatic symptom distress: A bifactor model approach. *Psychol Assess* 2016;28:109–15.
16 Cano-García FJ, Muñoz-Navarro R, Sesé Abad A, *et al*. Latent structure and factor invariance of somatic symptoms in the patient health questionnaire (PHQ-15). *J Affect Disord* 2020;261:21–9.
17 Page MJ, McKenzie JE, Bossuyt PM, *et al*. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 2021;372:71.
18 Rosendal M, Olde Hartman TC, Aamland A, *et al*. "Medically unexplained" symptoms and symptom disorders in primary care: prognosis-based recognition and classification. *BMC Fam Pract* 2017;18:18:18.:.
19 Lehmann M, Pohontsch NJ, Zimmermann T, *et al*. Diagnostic and treatment barriers to persistent somatic symptoms in primary care - representative survey with physicians. *BMC Fam Pract* 2021;22:60.
20 Kitselaar WM, van der Vaart R, van Tilborg-den Boeft M, *et al*. The general practitioners perspective regarding registration of persistent somatic symptoms in primary care: a survey. *BMC Fam Pract* 2021;22.
21 Marks EM, Hunter MS. Medically unexplained symptoms: an acceptable term? *Br J Pain* 2015;9:109–14.
22 Henningsen P, Zipfel S, Sattel H, *et al*. Management of functional somatic syndromes and bodily distress. *Psychother Psychosom* 2018;87:12–31.
23 Henningsen P, Jakobsen T, Schiltenwolf M, *et al*. Somatization revisited: diagnosis and perceived causes of common mental disorders. *J Nerv Ment Dis* 2005;193:85–92.
24 Rief W, Martin A, Rauh E, *et al*. Evaluation of general practitioners' training: how to manage patients with unexplained physical symptoms. *Psychosomatics* 2006;47:304–11.
25 Gendelman O, Amital H, Bar-On Y, *et al*. Time to diagnosis of fibromyalgia and factors associated with delayed diagnosis in primary care. *Best Practice & Research Clinical Rheumatology* 2018;32:489–99.
26 Berger A, Sadosky A, Dukes E, *et al*. Characteristics and patterns of healthcare utilization of patients with fibromyalgia in general practitioner settings in germany. *Curr Med Res Opin* 2008;24:2489–99.
27 Konnopka A, Schaefert R, Heinrich S, *et al*. Economics of medically unexplained symptoms: a systematic review of the literature. *Psychother Psychosom* 2012;81:265–75.
28 Zonneveld LNL, Sprangers MAG, Kooiman CG, *et al*. Patients with unexplained physical symptoms have poorer quality of life and higher costs than other patient groups: a cross-sectional study on burden. *BMC Health Serv Res* 2013;13:520.
29 Franks P, Clancy CM, Nutting PA. Gatekeeping revisited--protecting patients from overtreatment. *N Engl J Med* 1992;327:424–9.
30 Loudon I. The principle of referral: the gatekeeping role of the GP. *Br J Gen Pract* 2008;58:128–30.
31 Külekçioğlu S. Diagnostic difficulty, delayed diagnosis, and increased tendencies of surgical treatment in fibromyalgia syndrome. *Clin Rheumatol* 2022;41:831–7.
32 Kitselaar WM, Numans ME, Sutch SP, *et al*. Identifying persistent somatic symptoms in electronic health records: exploring multiple theory-driven methods of identification. *BMJ Open* 2021;11:e049907.
33 Kohlmann S, Löwe B, Shedden-Mora MC. Health care for persistent somatic symptoms across europe: A qualitative evaluation of the EURONET-SOMA expert discussion. *Front Psychiatry* 2018;9:646:646.:.
34 Henningsen P. Management of somatic symptom disorder. *Dialogues Clin Neurosci* 2018;20:23–31.
35 Kitselaar WM, van der Vaart R, Perschl J, *et al*. Predictors of persistent somatic symptoms in the general population: a systematic review of cohort studies. *Psychosom Med* 2023;85:71–8.
36 Hinz A, Ernst J, Glaesmer H, *et al*. Frequency of somatic symptoms in the general population: normative values for the patient health questionnaire-15 (PHQ-15). *J Psychosom Res* 2017;96:S0022-3999(16)30453-6:27–31.:.
37 Terluin B, van Marwijk HWJ, Adèr HJ, *et al*. The four-dimensional symptom questionnaire (4DSQ): A validation study of A multidimensional self-report questionnaire to assess distress, depression, anxiety and somatization. *BMC Psychiatry* 2006;6:34.
38 Toussaint A, Hüsing P, Kohlmann S, *et al*. Detecting DSM-5 somatic symptom disorder: criterion validity of the patient health questionnaire-15 (PHQ-15) and the somatic symptom scale-8 (SSS-8) in combination with the somatic symptom disorder - B criteria scale (SSD-12). *Psychol Med* 2020;50:324–33.
39 Kop R, Hoogendoorn M, Teije AT, *et al*. Predictive modeling of colorectal cancer using a dedicated pre-processing pipeline on routine electronic medical records. *Comput Biol Med* 2016;76:30–8.
40 Półchłopek O, Koning NR, Büchner FL, *et al*. Quantitative and temporal approach to utilising electronic medical records from

general practices in mental health prediction. *Comput Biol Med* 2020;125:103973.
41 Koning NR, Büchner FL, Vermeiren RRJM, *et al*. Identification of children at risk for mental health problems in primary care-development of a prediction model with routine health care data. *EClinicalMedicine* 2019;15:89–97.
42 ICPC | NHG. Available: https://www.nhg.org/themas/artikelen/icpc [Accessed 10 Nov 2020].
43 WCCfDS M. ATC index with ddds. Available: 2002.https://www.whocc.no/atc_ddd_index/ [Accessed 14 Nov 2020].
44 NHG. Available: https://referentiemodel.nhg.org/sites/default/files/NHG-Tabel [Accessed 20 Nov 2020].
45 STIZON - stichting informatievoorziening voor zorg en onderzoek. n.d. Available: https://www.stizon.nl/
46 WONCA. *ICPC- 2-R: international classification of primary care*. 2005.
47 Guidelines for ATC classification and DDD assignment 2013. Oslo. 2012. Available: https://www.whocc.no/filearchive/publications/1_2013guidelines.pdf
48 Zaki MJ. Spade: an efficient algorithm for mining frequent sequences. *Mach Learn* 2001;42(1/2):31–60.
49 McNeish DM. Using LASSO for predictor selection and to assuage overfitting: a method long overlooked in behavioral sciences. *Multivariate Behav Res* 2015;50:471–84.
50 Perlato A. Deal multicollinearity with lasso regression. 2019. Available: https://www.andreaperlato.com/mlpost/deal-multicollinearity-with-lasso-regression/ [Accessed 31 Mar 2022].
51 Sarraju A, Ward A, Chung S, *et al*. Machine learning approaches improve risk stratification for secondary cardiovascular disease prevention in multiethnic patients. *Open Heart* 2021;8:e001802.
52 Abidi L, Oenema A, van den Akker M, *et al*. Do general practitioners record alcohol abuse in the electronic medical records? A comparison of survey and medical record data. *Curr Med Res Opin* 2018;34:567–72.
53 Molinaro AM. Diagnostic tests: how to estimate the positive predictive value. *Neurooncol Pract* 2015;2:162–6.
54 Monden R, Rosmalen JGM, Wardenaar KJ, *et al*. Predictors of new onsets of irritable bowel syndrome, chronic fatigue syndrome and fibromyalgia: the lifelines study. *Psychol Med* 2022;52:112–20.
55 Schelde AB, Kornholt J. Validation studies in epidemiologic research: estimation of the positive predictive value. *J Clin Epidemiol* 2021;137:262–4.
56 Warren JW, Clauw DJ. Functional somatic syndromes: sensitivities and specificities of self-reports of physician diagnosis. *Psychosom Med* 2012;74:891–5.
57 Smith RC, Gardiner JC, Armatti S, *et al*. Screening for high utilizing somatizing patients using A prediction rule derived from the management information system of an HMO: A preliminary study. *Med Care* 2001;39:968–78.
58 Morriss R, Lindson N, Coupland C, *et al*. Estimating the prevalence of medically unexplained symptoms from primary care records. *Public Health* 2012;126:846–54.
59 Masters ET, Emir B, Mardekian J, *et al*. Identification of a potential fibromyalgia diagnosis using random forest modeling applied to electronic medical records. *JPR* 2015;288:277.
60 Pohontsch NJ, Zimmermann T, Jonas C, *et al*. Coding of medically unexplained symptoms and somatoform disorders by general practitioners-an exploratory focus group study. *BMC Fam Pract* 2018;19:129:129.:.
61 Jeffery DD, Bulathsinhala L, Kroc M, *et al*. Prevalence, health care utilization, and costs of fibromyalgia, irritable bowel, and chronic fatigue syndromes in the military health system, 2006-2010. *Mil Med* 2014;179:1021–9.
62 Masters ET, Mardekian J, Emir B, *et al*. Electronic medical record data to identify variables associated with a fibromyalgia diagnosis: importance of health care resource utilization. *JPR* 2015;8:131.