Promoting early recognition of persistent somatic symptoms in primary care

Kitselaar, W.M.

Chapter 6

Early Identification of Persistent Somatic Symptoms in Primary Care: data- and theory-driven predictive modelling based on electronic medical records of Dutch general practices

Willeke M. Kitselaar, Frederike L. Büchner, Rosalie van der Vaart, Stephen P. Sutch, Frank C. Bennis, Andrea W.M. Evers, Mattijs E. Numans

Based on: Kitselaar WM, Büchner FL, van der Vaart R, Sutch SP, Bennis FC, Evers AWM, Numans ME. 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;0:e066183. DOI:10.1136/ bmjopen-2022-066183

Abstract

Objective: The present study aimed to early identify patients with persistent somatic symptoms (PSS) in primary care by exploring 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 a first PSS registration in 2017-2018. Candidate predictors were selected 2-5 years prior to first registration of PSS and categorized 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 LASSO regression on 80% of the dataset. Derived models were internally validated on the remaining 20% of the dataset.

Results: All models had comparable predictive value (AUCs=.70-.72). Predictors are related to genital complaints, specific symptoms (e.g., digestive, fatigue, mood), health care utilization, 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.

Conclusions: This study shows that a simple clinical decision rule based on structured symptom/disease- or medication codes could possibly be an efficient way to support GPs in identifying patients in need of a different diagnostic or care approach. A fully 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.

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] as well as in patients with symptoms without well-established biomedical pathology.[1] PSS are not only burdensome to the patient,[7] but also greatly impact health care.[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 health care 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 utilization and costs.[26-28] Especially in many western countries, where general practitioners (GPs) serve as a gatekeeper for specialist health care.[29,30] To prevent unnecessary referrals and medicalization, 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 identification of PSS in primary care.[10,19] For example, diagnosis may be difficult due to 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 for 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, wide-spread 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 a 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 two-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 generalizable to clinical practice. Moreover, since we

cannot control variable collection, we are interested in how theory-driven variable selection performs compared to 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 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, onset date of PSS was determined according to the approach described below (see 'Outcome' section) within the period 1st of January 2017 until 31st of December 2018 (random 'onset' dates were selected for patients without PSS). Thereafter, candidate predictors were selected 2 to 7 years prior to the onset date (i.e., 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 (e.g. 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 datawarehouse 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 pseudonymized 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 December 31st of 2018, were excluded from the study.

Because we were interested in PSS onset prediction, patients who were registered with PSS before the 1st of 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 in the episode description; and /or (3) a score of ≥ 20 on the somatization subscale of the four-dimensional symptom questionnaire (4DSQ), registered in the lab-results. For a more detailed description of the selection criteria see [32].

Candidate predictors

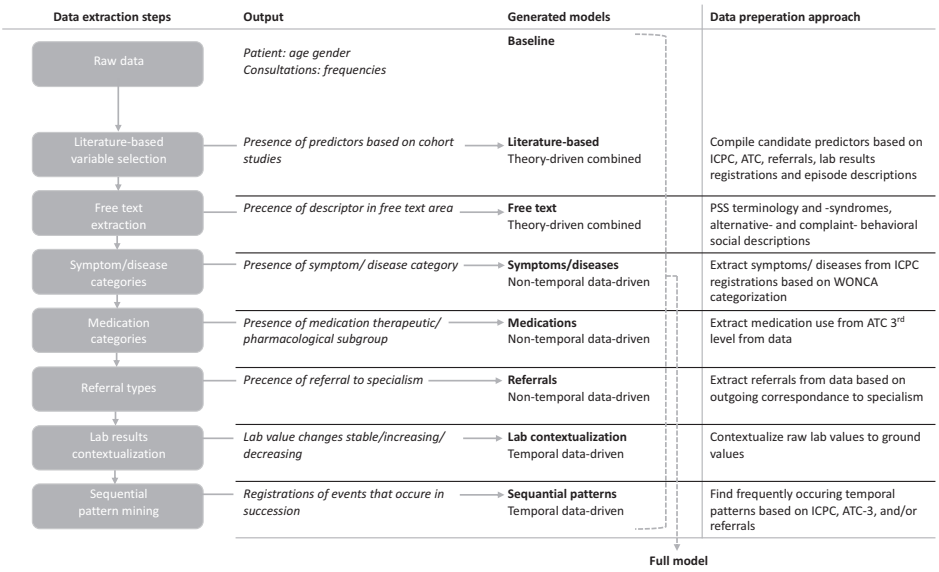
Different datasets were constructed with specific theory and data-driven candidate predictors of PSS in the ELAN data. Below a brief description of the predictor categories related to each dataset-based model will be given, see figure 1 for an overview of the data extraction steps and appendix A for a detailed overview of candidate predictors. Two distinct theory-driven datasets were operationalized; (1) literature-based risk factors of PSS, (see [35] for more detail) and; (2) frequencies of specific PSS-related terms and words in the free text with limited structured registration options (see appendix A). Data-driven datasets were divided into non-temporal and temporal data-driven datasets. The non-temporal datasets consist of dichotomized medical coding data (symptom/disease codes, medication codes, and referrals). The coded symptom/disease dataset was based on ICPC codes categorized into WONCA chapters and code categories.[46] The coded medication dataset was based on ATC codes reduced to 3rd level (to therapeutic/pharmacological subgroup).[47] The referral dataset was based on correspondences GPs have with other health care professionals.

The temporal approach consists of contextualized lab results and sequential patterns in medical coding data. Due to the high number of different lab results and inconsistent availability, using reference values for this study was not feasible. Contextualization of

lab results provide 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 patient has multiple registrations on one day these will be separated and combined with possible subsequent registrations (e.g., patient X has the following registrations on date Y: fatigue, abdominal pain, anti-constipation drug and date Z: physiotherapy, this will result in 3 patterns for patient X: (1) fatigue → physiotherapy; (2) abdominal pain → physiotherapy; (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 (i.e., prevalence of the pattern in de dataset). Please see [48] for a more detailed description of the SPADE algorithm.

Figure 1. Diagram showing the data extraction steps for each constructed model.



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 5-fold cross-validation, with hyperparameter tuning, was performed on the training set. For each unique model (i.e., literature-review, free text, coded symptom/diseases, coded medications, referrals, contextualization of lab results, and sequential patterns) and all combined models (i.e., theory-driven, data-driven nontemporal, data-driven temporal, and full model), near zero-variance candidate

predictors were removed (see appendix B 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 and the area under the ROC curve (AUC) was calculated. All data was prepared and analysed using R version 4.0. For the final modelling, the caret-package was used.

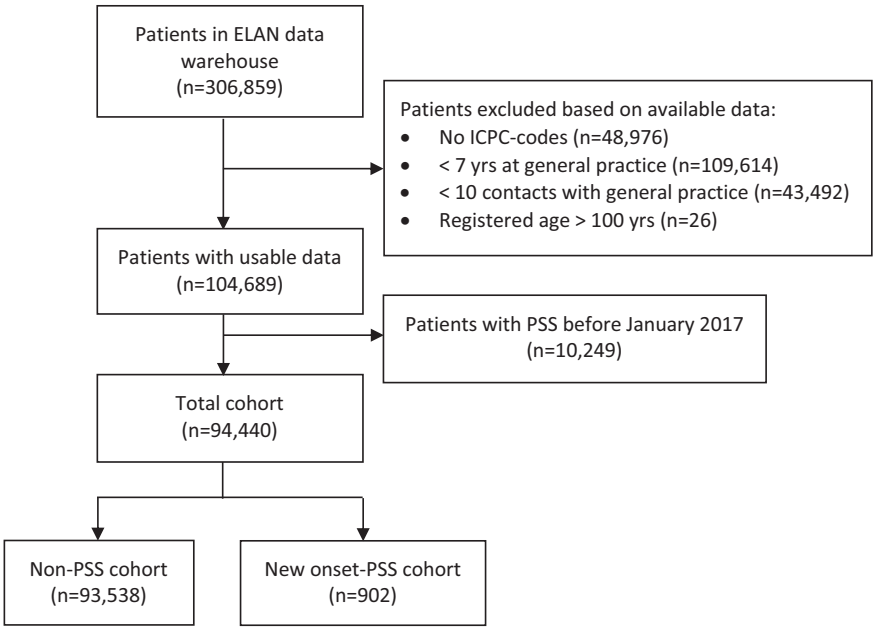
Final model evaluation

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 odds ratios (ORs). To verify the stability of the predictor estimates, frequencies of estimates receiving non-zero values were calculated across 1000 bootstrap samples.

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 January 1st and therefore also excluded from the study. As a result, 94,440 patients were included in the final analysis (figure 2).

Figure 2. flow chart patient inclusion ELAN study cohort



As shown in table 1, 0.9% (n=902) of patients in the ELAN cohort had new-onset PSS. Compared to the total cohort, patients with PSS are more likely to be female (69.0% vs. 52.9%), are generally younger (52.6±14.4 vs. 57.2±15.4) and have higher consultation frequency (8.7±7.3 vs. 6.3±5.8). Moreover, patients with PSS are more likely to have a mental health disorder (60.3% vs. 46.8%) while the likelihood of a physical disorder does not differ (64.6% vs. 63.6%, p = .87). The patients with new-onset PSS in the training and test set differ on baseline variable female (68.3% vs. 72.2%). Post-hoc evaluation revealed that patients with PSS in the training and test set 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).

Table 1. Patient characteristics

	Total cohort	PSS		
	Full dataset	**Full dataset**	**Training**	**Test**
n (%)	94440 (100.00)	902 (0.9)	772 (0.9)	180 (0.9)
Female, n (%)	49998 (52.9)	623 (69.0)	493 (68.3)	130 (72.2)
Age, mean (s.d.)	57.2 (15.4)	52.6 (14.4)	52.9 (14.5)	51.3 (13.7)
Consultations, mean (s.d.)	6.3 (5.8)	8.7 (7.3)	7.44 (6.3)	7.2 (5.5)
Urbanization level, n (%)				
Urban area	45567 (48.2)	404 (44.8)	326 (45.2)	78 (43.3)
Sub-urban area	43296 (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	67215 (71.2)	622 (69.0)	494 (68.4)	128 (71.1)
Physical comorbidity, n (%)	60019 (63.6)	583 (64.6)	470 (65.1)	113 (62.8)
Mental comorbidity, n (%)	44292 (46.9)	544 (60.3)	430 (59.6)	114 (63.3)

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 .68 for the baseline model to .72 for the full model. From the separate models, all models preformed equally well, based on an approximate AUC .70. Using the optimal cut-off selection (i.e., 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 the appendixes A-C for more details on the model contents).

Table 2. Prediction models based on LASSO logistic regression analysis

			TRAINING	TEST		
			AUC	**Sensitivity**	**Specificity**	**AUC**
Theory-driven		**Baseline model** [a]	**.66**	**.73**	**.54**	**.68**
		Literature-based [b, c]	.70	.61	.68	.71
		Free text [b, d]	.68	.70	.56	.71
		Combined [a]	**.69**	**.73**	**.60**	**.71**
Non-temporal Data-driven		Symptoms/diseases [b, e]	.68	.72	.57	.70
		Medications [b, f]	.69	.76	.58	.70
		Referrals [b, g]	.66	.71	.55	.69
		Combined [b]	**.70**	**.57**	**.72**	**.71**
Temporal Data-		Lab contextualization [b, h]	.67	.73	.58	.70
		Sequential patterns [b, i]	.66	.83	.43	.69
		Combined [b]	**.68**	**.73**	**.58**	**.70**
		Full model [b, j]	**.70**	**.72**	**.60**	**.72**

[a] Gender, age, consultation frequency; [b] includes baseline model; [c] Variables selected based on literature search of risk factors in the general population; [d] Word search through free journal text; [e] ICPC-codes categorized according to the WONCA categorization (dichotomized); [f] ATC-3: therapeutic/pharmacological subgroup (dichotomized); [g] Outgoing correspondence to medical specialists (dichotomized); [h] Relative grounded lab-results (stable, increase, decrease; dichotomized); [I] Order of ICPC, ATC and referrals over time, patterns identified with the SPADE algorithm (see appendix C); [j] All available candidate predictors combined; For a detailed description of the models, see appendix A

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 contextualization (n=3), and sequential patterns (n=1). From the baseline predictors, age decreased (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).

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

Predictors	Total cohort % or mean (s.d.)	PSS-cohort % or mean (s.d.)	Odds ratio	%[a]
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) [b]	1.1	3.1	1.17	60.8
Medications [c]	2.0 (1.4)	2.5 (1.6)	1.12	94.7
Number of imaging referrals [d]	0.09 (0.09)	0.1 (0.1)	1.10	96.1
Fatigue [e]	20.5	31.2	1.04	47.5
Mood disorder [f]	14.6	23.6	1.03	47.7
Number of pain sites [g]	0.3 (0.6)	0.5 (0.7)	1.02	63.7
Headache [h]	19.8	32.6	1.02	44.8
Number of ICPC-codes [i]	2.6 (1.5)	3.3 (1.7)	1.004	13.5
Free text (theory driven)				
Complaint description [j]	0.7 (1.0)	1.3 (1.6)	1.12	99.3
PSS terminology [k]	0.06 (0.15)	0.11 (0.21)	1.04	83.6
Symptom/disease codes (non-temporal data-driven)				
Neurological disorder [l]	18.1	27.3	1.11	77.9
Digestive symptoms [m]	50.4	65.5	1.07	66.7
Female genital symptoms [n]	28.8	46.6	1.07	53.0
Female genital infection [o]	8.3	15.9	1.04	48.9
Medication codes (non-temporal data-driven)				
Capillary stabilizers [p]	0.1	0.7	1.47	57.6
Selective CA+ blockers [q]	10.6	6.3	0.93	58.0
Topical contraceptives [r]	5.5	10.5	1.06	58.8
Lipid modifier [s]	21.4	15.6	0.95	54.2
Nasal spray, topical [t]	40.1	51.7	1.02	51.1
Anti-constipation drug [u]	28.4	40.1	1.02	52.1
Eyedrops, topical [v]	16.2	22.3	1.01	47.3
Anti-thrombotic agents [w]	20.8	16.0	0.999	41.0

Table 3. Predictors of PSS obtained from full model LASSO logistic regression analysis (continued)

Predictors	Total cohort % or mean (s.d.)	PSS-cohort % or mean (s.d.)	Odds ratio	%[a]
Referrals (non-temporal data-driven)				
Physiotherapy [x]	30.2	39.5	1.01	43.6
Lab contextualization (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

[a] Frequency of estimates having non-zero values across 1000 bootstrap samples; [b] ICPC-codes: X04, P08.02; [c] Frequency based on full ATC codes; [d] Rontgen or echography; [e] ICPC-code: A04; [f] ICPC codes: P03, P73, P73.02, P76 and ATC codes: N06A, N05AN, D11AX04; [g] Number of pain-related ICPC codes; [h] ICPC codes: N01, N02, N89, N90, R09; [i] all unique ICPC codes; [j] fatigue, dizziness, back pain (see appendix A for full list); [k] e.g., somatization or a-specific symptoms (see appendix A for full list); [l] ICPC: N86-99; [m] ICPC codes: D01-29; [n] ICPC codes: X01-29; [o] ICPC codes: X70-74 and X90-92; [p] ATC4-codes: C05C; [q] ATC4 codes: C08C; [r] ATC4 codes: G02B; [s] ATC4 codes: C10A; [t] ATC4 codes: R01A; [u] ATC4 codes: A06A; [v] ATC4 codes: S01X; [w] ATC4 codes: B01A; [x] Correspondence with physiotherapy.

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 two years prior to index date. Model performance based on specific predictor generation approaches do not differ greatly. Therefore, the use of the simplest approach may be most desirable. Based on the full model (including all candidate 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 (e.g., digestive issues, fatigue, mood disorders, and headache), health care utilization (e.g., number of medications or imaging, referrals, or physiotherapy), and number of complaints (e.g., 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.[52] To our knowledge, we included a wider range of predictors than previous studies, and these are clinically relevant and generalizable 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 (i.e., stable lab results: lymphocytes and thyroid). On the other hand, the use of routine care data may also limit the generalizability of the predictors to the general population since registration depend 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.[32,41,53] 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. The aim was to enable data-driven selection and not rely on GP diagnosis, since research indicates that PSS are often missed by physicians.[54] Data-driven selection would enhance re-usability of routine care data.

To our knowledge this is the first cohort study to predict PSS two 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 their use of questionnaires or physician dependent diagnoses.[55-57] A recent study based on the ELAN datawarehouse with a non-biomedical outcome showed similar predictive value,[41] which could mean that routine primary care data has 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 sub-types.[52,57] Monden et al.,[54] reported notably higher odds ratios, which may be related to less available confounding variables and/or to active data collection resulting in access to multidomain (i.e., more complete social and psychological) data. This is in line with studies showing that GPs are less likely to report social and psychological factors [19,20,58] and a recent systematic review demonstrating the importance of using multidomain data.[45] Lastly, in contrast to a body of evidence,[57,59,60] our LASSO regression of the full model did not indicate that

consultation frequency predicts PSS. Since consultation frequency was predictive in most sub-models, 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 indicates that it cannot be used without additional screening. Relatively simple ICPC/ATC-based models can assist in distinguishing between PSS and well-established biomedical problems. Predictive value 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 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 contextualization did not enhance model performance, the former implies that other machine learning techniques (e.g., text mining) should be further explored. Especially because of the fair performance of the free text-based model, for which in the present study only limited free text is utilized.

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 (e.g., 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 (i.e., PSS) using validated screening tools (e.g., 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, combing 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.

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.

References

1 Kop WJ, Toussaint A, Mols F, Löwe B. 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.

4 Katon W, Lin EHB, Kroenke K. The association of depression and anxiety with medical symptom burden in patients with chronic medical illness. Gen Hosp Psychiatry 2007; 29: 147–55.

5 Grassi L, Caruso R, Nanni MG. Somatization and somatic symptom presentation in cancer: A neglected area. International Review of Psychiatry 2013; 25: 41–51.

6 Kohlmann S, Gierk B, Hummelgen M, Blankenberg S, Lowe B. 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, Löwe B, Rief W. Functional somatic disorders: Discussion paper for a new common classification for research and clinical use. BMC Med 2020; 18: 1–7.

9 Haller H, Cramer H, Lauche R, Dobos G. 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, Lowe B. 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 Lowe 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, Rist F, Nater UM. 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.
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: 1–9.
19 Lehmann M, Pohontsch NJ, Zimmermann T, Scherer M, Löwe B. Diagnostic and treatment barriers to persistent somatic symptoms in primary care - representative survey with physicians. BMC Fam Pract 2021; 22.
20 Kitselaar WM, van der Vaart R, van Tilborg-den Boeft M, Vos HMM, Numans ME, Evers AWM. 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, Creed F. Management of Functional Somatic Syndromes and Bodily Distress. Psychother Psychosom 2018; 87: 12–31.
23 Henningsen P, Jakobsen T, Schiltenwolf M, Weiss MG. 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, Zech T, Bender A. 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 Pract Res Clin Rheumatol 2018; 32: 489–99.
26 Berger A, Sadosky A, Dukes E, Martin S, Edelsberg J, Oster G. Characteristics and patterns of healthcare utilization of patients with fibromyalgia in general practitioner settings in Germany. https://doi.org/101185/03007990802316550 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 LN, Sprangers MA, Kooiman CG, van 't Spijker A, Busschbach JJ. 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. New England Journal of Medicine. 1992; 327: 424–9.
30 Loudon I. The principle of referral: The gatekeeping role of the GP. British Journal of General Practice 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, Faiq A, Evers AW, van der Vaart R. 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.

34 Henningsen P. Management of somatic symptom disorder. Dialogues Clin Neurosci 2018; 20: 23-+.

35 Kitselaar WM, van der Vaart R, Perschl J, Numans ME, Evers AWM. Predictors of Persistent Somatic Symptoms in the General Population: A Systematic Review of Cohort Studies. Psychosomatic Medicine 85(1):71-8, January 2023.

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: 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.

38 Toussaint A, Hüsing P, Kohlmann S, Löwe B. 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 A ten, 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, Crone MR, Numans ME, Hoogendoorn M. 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, Crone MR, Numans ME. 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.

43 WCCfDS M. ATC index with DDDs. 2002.

44 NHG. Tabel 12 - soort derden.

45 STIZON - Stichting Informatievoorziening voor Zorg en Onderzoek.

46 WONCA. ICPC- 2-R: International Classification of Primary Care. 2005.

47 Guidelines for ATC classification and DDD assignment 2013. Oslo, 2012.

48 Zaki MJ. SPADE: An Efficient Algorithm for Mining Frequent Sequences. Machine Learning 2001 42:1 2001; 42: 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.

51 Sarraju A, Ward A, Chung S, Li J, Scheinker D, Rodríguez F. Machine learning approaches improve risk stratification for secondary cardiovascular disease prevention in multiethnic patients. Open Heart 2021; 8: e001802.

52 Monden R, Rosmalen JGM, Wardenaar KJ, Creed F. Predictors of new onsets of irritable bowel syndrome, chronic fatigue syndrome and fibromyalgia: The lifelines study. Psychol Med 2020; : 1–9.

53 Abidi L, Oenema A, van den Akker M, van de Mheen D. 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.

54 Warren JW, Clauw DJ. Functional somatic syndromes: Sensitivities and specificities of self-reports of physician diagnosis. Psychosom Med 2012; 74: 891–5.

55 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.

56 Morriss R, Lindson N, Coupland C, Dex G, Avery A. Estimating the prevalence of medically unexplained symptoms from primary care records. Public Health 2012; 126: 846–54.

57 Emir B, Masters ET, Mardekian J, Clair A, Kuhn M, Silverman SL. Identification of a potential fibromyalgia diagnosis using random forest modeling applied to electronic medical records. J Pain Res 2015; 8: 288.

58 Pohontsch NJ, Zimmermann T, Jonas C, Lehmann M, Löwe B, Scherer M. Coding of medically unexplained symptoms and somatoform disorders by general practitioners – an exploratory focus group study. BMC Fam Pract 2018; 19: 129.

59 Jeffery DD, Bulathsinhala L, Kroc M, Dorris J. 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.

60 Masters ET, Mardekian J, Emir B, Kuhn M, Silverman SL. electronic medical record data to identify variables associated with a fibromyalgia diagnosis: importance of health care resource utilization. J Pain Res 2015; 8: 131–8.

Appendixes

Appendix A. Predictors derived from models based on LASSO regression

Baseline model	Gender, age, consultation frequency				
Literature-based*	Urbanization, deprived neighbourhood, frequency of referral to imaging, frequency of referral to psychology, frequency of referral to alternative medicine, frequency of referral to ER, frequency of referral for secondary care, frequency of referral to primary care, frequency of referral for laboratory tests, variation in medication prescription (full length ATC), variation in ICPC codes, anxiety, number of pain symptoms, arterial pathology, asthma, atopy, burn injury, BMI, burn, CTS, birth, cholesterol, chronic illness, chronic kidney disease, chronic sinus, chronic stress, conduct problems, COPD, coronary artery disease, dementia, diabetes, diffuse pain, dizziness, dyslipidaemia, dyspareunia, dyspepsia, employment, family history of disease, fatigue, gastrointestinal symptoms, headache, health anxiety, heart failure, hormonal medication, hypertension, hyperthyroidism, vaccinations, infections, life events, liver disease, malignant neoplasm, marital status, memory problems, mental health, menstrual disorders, Meniere disease, mood disorders, musculoskeletal disease, neuritis, neuropathic pain, non-specific complaints, osteoporosis, pain medications, psoriasis, restless-leg syndrome, rheumatism, SES, abuse, sleep apnoea, sleep disorder, smoking, somatic symptoms, specific pain, stroke, teeth grinding, traffic accident, traumatic brain injury.				
Free text*	**PSS TERMINOLOGY:** MUPS, somatization, psychosomatic, unexplained, functional complaints, central sensitization, somatization disorder, somatically unexplained, complaints, somatoform, misunderstood, complaints, neurasthenia functional, barriers, vague complaints, vague pain, non-specific,	**PSS SYNDROMES:** Fibromyalgia, spastic colon, irritable bowel, gut syndrome, IBS, CFS, chronic fatigue, ME/CFS, tinnitus, facial pain, vulvodynia, restless legs syndrome, bladder syndrome, bladder pain syndrome, interstitial cystitis, unstable bladder, tension headache,	**ALTERNATIVE DESCRIPTION**: Stagnant, recovery, persistent, working hypothesis, no abnormalities, impediments, meaningless, pain experience, illness anxiety, negative thoughts, fear of movement, experiences, to experience, complaint-contingent approach, sensitive, load capacity, explanatory model.	**COMPLAINT DESCRIPTION**: Dizziness, fatigue, concentration, tension, stress-related, generalized/staggering pain, hypermobile, low back, SI pain, lumbago, backpain, pseudo-radicular, tendinosis, muscle-joint pain, musculoskeletal system, memory problems, headache, tingling, dispirited, rebellious,	**BEHAVIOR-SOCIAL:** Avoidance behaviour, absenteeism, surroundings, social problems, functioning, culture, tensions, traumatic event, abuse, addiction, violence, domestic.

Appendix A. Predictors derived from models based on LASSO regression (continued)

Free text* (continued)	**PSS TERMINOLOGY:** reactive complaints, unexplained complaints, stress complaints, stress complaints.	**PSS SYNDROMES:** pain syndrome.	**ALTERNATIVE DESCRIPTION**:	**COMPLAINT DESCRIPTION**: desperate, depressed, sleep, nauseous, shiver, anxiety symptoms, angry, anxious, emotional, dejected, worry, listless, upset stomach, on chest, neck pain, itch, sad, gloom	**BEHAVIOR-SOCIAL:**
Symptoms/diseases*	WONCA categorized ICPC-codes: general/unspecified congenital anomalies (A90), general/unspecified infections (A70-78), general/unspecified injuries (A80-89), general/unspecified other diagnoses (A91-99), general/unspecified non-specific symptoms (A01, A05, A20, A28-29), general/unspecified specific symptoms (A02-06, A08-09, A12), blood/immune Infections (B70-71), blood/immune other diagnoses (B80-99), blood/immune symptoms (B02-29), digestive infections (D70-73), digestive neoplasms (D74-78), digestive other diagnoses (D82-99), digestive symptoms (D01-29), eye infections (F70-73), eye injuries (F75-79), eye other diagnoses (F82-99), eye symptoms (F01-29), ear infections (H70-74), ear injuries (76-79), ear other diagnoses (H81-99), ear symptoms(H01-29), cardiovascular congenital (K73), cardiovascular other diagnoses (K74-99), cardiovascular symptoms (K01-29) musculoskeletal injuries (L72-81), musculoskeletal other diagnoses (L83-95), musculoskeletal non-specific symptoms (L18-20, L28-29), musculoskeletal specific symptoms (L01-17), neurological neoplasms (N74-76), neurological other diagnoses (N86-99), neurological symptoms (N04, N06-08), psychological other diagnoses (P70-99, T06), psychological symptoms (P01-29), respiratory infections (R70-83), respiratory injuries (R87-88), respiratory neoplasms (R84-86), respiratory other diagnoses (R90-99), respiratory symptoms (R01-29), skin congenital (S81-83), skin infections (S03, S09-11, S84, S95), skin injuries (S12-19), skin neoplasms (S77-80), skin other diagnoses (S84-94, S96-99), skin symptoms (S01-29), endocrine/metabolic other diagnoses (T81-99), endocrine/metabolic symptoms (T01-29), urological other diagnoses (U88-99), urological symptoms (U01-29), family planning other diagnoses (W77-99), family planning symptoms (W01-29), female genital infections (X70-74, X90-91), female genital neoplasms (X75-81), female genital other diagnoses (X84-89, X99), female genital symptoms (X01-29), male genital congenital (Y81-84), male genital infections (Y70-76), male genital other diagnoses (Y85-99), male genital symptoms (Y01-29), social symptoms (Z01-29).				
Medications*	ATC 3rd level: therapeutic/pharmacological subgroup				

Appendix A. Predictors derived from models based on LASSO regression (continued)

Referrals*	Acupuncture, allergology, anaesthetics, autography, cardiology, surgery, cytology, dermatology, dietarian, primary care psychologist, endocrinology, physiotherapy, mental health care, gynaecology, haptomology, internal medicine, ear-nose-throat specialist, laboratory testing, pneumology, gastroenterology, medical microbiology, neurology, optomologist, orthopedy, plastic surgery, pain relief centre, podiatry, psychology, psychotherapy, radio therapy, rheumatology, rehabilitation centre, Rontgen, emergency care, urologist.
Lab contextualization*	Bilirubin, cholesterol, creatine, CRP/BSE, glucose, granulocyte, HbA1c, haemoglobin, minerals, monocytes, neutrophiles, PH (urine), systolic blood pressure, thyroid function, transaminase, vitamin B (excl. B12), vitamin B12, vitamin D, weight/BMI
Sequential patterns*	*3-level patterns*: Antibacterial drugs (systemic; ATC-code: J01) >> secondary care referrals, analgesic drugs (ATC-code: N02) >> secondary care referral *2-level patterns*: hypertensive heart disease >> secondary care referral, specific musculoskeletal symptoms (ICPC-codes: L01-17) >> secondary care referral, Rontgen referral >> secondary care referral, hypertensive heart disease >> Rontgen referral, specific musculoskeletal symptoms (ICPC-codes: L01-17) >> Rontgen referral, antibacterial drugs (systemic; ATC: J01) >> specific musculoskeletal symptoms (ICPC-codes: L01-17), hypertensive heart disease >> specific musculoskeletal symptoms (ICPC-codes: L01-17), Rontgen referral >> specific musculoskeletal symptoms (ICPC-codes: L01-17), secondary care referral >> musculoskeletal disease (ICPC-codes: L83-95, L98-99), specific musculoskeletal symptoms (ICPC-codes: L01-17) > analgesic drugs (ATC-code: N02). *1-level patterns*: General and unspecified disease (ICPC-codes: A91-99), fatigue (ICPC-code: A04), no disease (ICPC-code: A97), abdominal symptoms (ICPC-codes: D01-29), peripheral osteoarthritis (ICPC-codes: L89-91), drugs for acid related disorders (ATC-code: A02), drugs for constipation (ATC-code: A06), vitamin preparations (ATC-code: A11), antithrombotic agents (ATC-code: B01), dermatological corticosteroids (ATC-codes: D07), antibacterial drugs (systemic; ATC: J01), analgesic drugs (ATC-code: N02), drugs for obstructive airway diseases (ATC-codes: R03), cough and cold preparations (ATC-codes: R05), ophthalmological drugs (ATC-codes: S01), acute unitary infection (ICPC-codes: U70-72), cancer (ICPC-codes: A79, B72-73, D74-77, L71, N74, R84-85, S77, T71, U75-77, W72, X75-77, Y77-78), chronic abdominal pain (ICPC-codes: D01-02, D04, D06, Y02) dizziness (ICPC-codes: H82, N17), eye symptoms, eye diseases (ICPC-codes: F83-84, F92-94), hypertensive heart disease (ICPC-codes: K86-87), cardiovascular other diagnoses (ICPC-codes: K74-99), cardiovascular symptoms (ICPC-codes: K01-29), musculoskeletal injuries (ICPC-codes: L72-81), musculoskeletal other diagnoses (ICPC-codes: L83-95, L98-99), specific musculoskeletal symptoms (ICPC-codes: L01-17), neck and shoulder symptoms (ICPC-codes: L01, L08), psychological symptoms (ICPC-codes: P01-29), respiratory symptoms (ICPC-codes: R01-29), Rontgen referral, skin other diagnoses (ICPC-codes:), skin symptoms (ICPC-codes: S01-29), secondary care referral, vitamin deficiency (ICPC-codes: T91), infections upper respiratory tract (ICPC-codes: A77, R72, R74-76), urological symptoms (ICPC-codes: U01-29), female genital symptoms (ICPC-codes: X01-29).

* Near zero variance and high-correlating variables removed

Appendix B. Number of variables by dataset and source table

Datasets	Source table(s)	n
Baseline	Patient	3
Symptoms/diseases	Journal and episode	96
Medications	Medication	176
Referrals	Correspondence	51
Literature review	Patient, journal, episode, lab, correspondence, medication	92
Free text	Journal	8
Lab contextualization	Lab results	76
Sequential patterns	Journal, episode, lab, correspondence, medication	57
Full model	Patient, journal, episode, lab, correspondence, medication	545

Appendix C. Patterns derived from the SPADE algorithm and subsequent LASSO regression for the sequential patterns model

Sequences	support (difference)	Odds ratio
Rontgen referral	0.077	1.08
Female genital symptom [a]	0.043	1.03
Hypertension [b]	0.036	0.97
Fatigue[c]	0.025	1.02
Antibacterials for systemic use >> specialist care referral	0.012	1.02
Antibacterials for systemic use >> specific musculoskeletal symptoms [d]	0.011	0.98
Drugs for constipation (A06)	0.274	1.00
Cardiovascular diagnosis [e]	0.031	1.00
Neck and back complaints [f]	0.036	1.00

[a] ICPC-codes X01-X29; [b] ICPC-codes K86 or K87; [c] ICPC-code A04; [d] ICPC-codes L01, L02, L03, L04, L05, L06, L07, L08, L09, L10, L11, L12, L13, L14, L15, L16, L17, L17.01; [e] ICPC codes K74-K99 (excl. K86 and K87); [f] ICPC codes L01, L02, L03, L83, L84, or L86