PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Early identification of persistent somatic symptoms in primary care: data- and theory-driven predictive modelling based on electronic medical records of Dutch general practices
AUTHORS	Kitselaar, Willeke; Büchner, Frederike; van der Vaart, Rosalie; Sutch, Stephen; Bennis, Frank; Evers, Andrea; Numans, Mattijs

VERSION 1 – REVIEW

REVIEWER	Schneider, Gudrun University Hospital Münster, Department of Psychosomatic Medicine and Psychotherapy
REVIEW RETURNED	29-Jul-2022

GENERAL COMMENTS	This is an interesting manuscript on an important topic. A strength ist the big number of GPs and patients included in this study. I only have some minor remarks. In the abstract there are abbreviations which are not explained.

REVIEWER	Silverberg, Noah The University of British Columbia
REVIEW RETURNED	28-Aug-2022

GENERAL COMMENTS	This study aims to predict which patients are at risk of developing PSS. They argue that patient-report screening tools are not feasible in a primary care setting, though it may be possible to develop prognostic models based on routinely collected clinical data. They used multiple types of predictors, including some novel and sophisticated approaches (e.g., derived from machine learning algorithms). Predictors were selected with LASSO and models evaluated with ROC, with training and test sets. The authors found that all prognostic models performed similarly, so "the use of the simplest approach may be most desirable." None of their prognostic models performed well (AUC>0.8), so they conclude that patient-report screening tools are necessary after all, bringing the paper full circle. Demonstrating that accurate prediction of new-onset PSS is not possible with routinely collected clinical data is still worthwhile. The authors should provide a rationale for why predicting the future onset of PSS is important. If a patient is flagged as being at elevated risk of developing PSS in 1-2 years, what is the primary care physician supposed to do with that information? The main findings hinge on their definition of PSS, which is "based on previous research" but should be explained more fully in the

	present paper. The authors acknowledged that their PSS case ascertainment method was “suboptimal” because incidence rates were lower than expected, but I don’t see reference 7-year incidence rates reported, only prevalence rates. To explain their lower than expected incidence rate, they offer “we cannot be certain that symptoms are not fully explained by a medical disorder.” This could explain overestimating, not underestimating, incidence. We are still left without any possible explanations for the low rate of PSS. It should also be noted that the low incidence of PSS is even more surprising given that only patients with at least 10 prior clinic contacts were eligible for this study. More than half of patients with PSS had a physical comorbidity. How can the authors be sure that the functional somatic symptoms were not fully accounted for by the physical comorbidities? Again, this gets at the PSS case ascertainment method. If the primary outcome is not truly PSS, that would explain some of the surprising null findings (number of clinic visits not associated with PSS). The authors should consider re-running their prognostic models in just the patients without physical comorbidities. Patients who already had PSS at the beginning of the observation were excluded. Wouldn’t this bias the sample to younger patients and those with relatively mild, transitory PSS? Minor concerns: -Please state rationale for why these cut-offs were chosen (1) registered at least 7 years at the same general practice, (2) >10 clinic visits, and (3) 2 year prediction interval -“While most of these symptoms are self-limiting,…” (page 4, line 13). What follows does not seem connected to this opening statement. -“Lastly, some physicians refrain from using terms beyond the biomedical domain for somatic symptoms” (page 4, line 27). Not sure what this means. -46.8% of the group without PSS had a mental health disorder. This seems very high. Was this an expected finding, consistent with known epidemiology? -The finding that stable lab test results were associated with onset of PSS is interesting and could be elaborate on -The Patient and public involvement statement requires elaboration. Who did the authors consult with, when, and about what. What was the outcome of the consultations? -Please add 95% confidence intervals to Table 3

REVIEWER	Kohlmann, Sebastian University Medical Center Hamburg-Eppendorf, Department of Psychosomatic Medicine and Psychotherapy
REVIEW RETURNED	29-Aug-2022

GENERAL COMMENTS	Short summary: The manuscript reports on a novel and interesting approach to early identify patients with persistent somatic symptoms (PSS) in primary care. The registry-based dataset is large and the analysis is complex. The results indicate that GP-based data can moderately contribute to predict the onset of PSS. The authors conclude that clinical decision rule based on structured symptom/disease- or medication codes can efficiently identify patients with PSS.

	General comment: The manuscript is an important contribution to the field of PSS and clinically relevant. In most parts it is well-written (introduction and discussion); in other parts the manuscript would benefit from presenting more information and illustration of the analysis (methods, results). The analysis appears to be sound but should be reviewed by expert in the field of machine learning. Strengths of the manuscript include its large dataset, the novel approach to identify PSS and the complex analysis. The main limitation, in my view, is that no patient-reported data was used in the prediction models. Additionally, I would slightly disagree with the authors' conclusion that routine primary care can be used to early predict PSS. The AUC values of the presented models can be judged as moderate (only). The results indicate that the early detection of individuals can be enhanced with data from primary but still there is large room for improvement. Below, I have made some suggestions to improve the manuscript. Major strengths: Innovation and relevant research idea Large dataset Major weaknesses: No patient reported outcomes in the data set Abstract: The readability of the abstract needs to be improved. The information is very hard to follow: there are many abbreviations and the analysis and the dataset is complex. The conclusion is not correct and should be formulated more cautiously. Introduction: The introduction is well written. The following papers may add some knowledge to recent developments in screening questionnaires and the current health care situation: Toussaint, A., Hüsing, P., Kohlmann, S., & Löwe, B. (2020). 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). Psychological Medicine, 50(2), 324-333. doi:10.1017/S003329171900014X Kohlmann, S., Löwe, B., & Shedden-Mora, M. C. (2018). Health care for persistent somatic symptoms across Europe: a qualitative evaluation of the EURONET-SOMA expert discussion. Frontiers in psychiatry, 9, 646. Methods: Why was having 10 or less contacts with general practice an exclusion criterion?

	With respect to the SPADE algorithm: Why was a “1% difference” chosen as relevant? Appendix S2 is not referenced in the text. Title of appendix 2 should be more precise: What is statistically meant with “most important patterns”? A figure that illustrates the regression model(s) with outcomes and statistical assumptions could improve understandability of the method section. Results: The sample description should include statistics on the differences between patients with and without PSS. Positive and negative predictive values should be mentioned in the text and table 2 Confidence intervals should be added with respect to the final predictor set (text and table 3). Are there any indicators to judge the validity of the model? The results would benefit from presenting a sensitivity analysis. E.g., What would the models’ properties be like if the outcomes were analyzed separately (ICPC-codes for PSS-syndromes vs. PSS-umbrella terms vs. 4DSQ? Another approach to show the validity of this diagnostic approach would be to show that the model’s prediction is comparable to established risk models to predict the course of somatic diseases, e.g. the “EURO Score” for coronary heart disease. Discussion: The discussion section is well-written and presents strengths and limitations. With respect to the main results, the authors’ interpretation should be done more cautiously: “Our study shows how routine primary care data can be used as a source that enables early prediction of PSS.” The AUC of the best model can be judged as moderate; specificity and sensitivity are not optimal to identify and rule out cases. Still, the results are important as this machine learning approach could be enriched with data from patients (PROMS) which might lead to better predictions. The authors should elaborate on the limitation that no PROMS were used and give further outlook with respect to existing studies on the criterion validity of available screeners (SSS-8, SSD-12, PHQ-15, DSQ-4, BDS, etc.).

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1

Dr. Gudrun Schneider, University Hospital Münster

Comments to the Author:

This is an interesting manuscript on an important topic. A strength ist the big number of GPs and patients included in this study.
I only have some minor remarks.

Thank you for your evaluation.

1. In the abstract there are abbreviations which are not explained.

Thank you for this comment. I have adjusted the abstract accordingly.

Reviewer: 1
Competing interests of Reviewer: none

****

Reviewer: 2

Dr. Noah Silverberg, The University of British Columbia

Comments to the Author:

2. This study aims to predict which patients are at risk of developing PSS. They argue that patient-report screening tools are not feasible in a primary care setting, though it may be possible to develop prognostic models based on routinely collected clinical data. They used multiple types of predictors, including some novel and sophisticated approaches (e.g., derived from machine learning algorithms). Predictors were selected with LASSO and models evaluated with ROC, with training and test sets. The authors found that all prognostic models performed similarly, so "the use of the simplest approach may be most desirable." None of their prognostic models performed well (AUC>0.8), so they conclude that patient-report screening tools are necessary after all, bringing the paper full circle. Demonstrating that accurate prediction of new-onset PSS is not possible with routinely collected clinical data is still worthwhile.

Thank you for this evaluation.

3. The authors should provide a rationale for why predicting the future onset of PSS is important. If a patient is flagged as being at elevated risk of developing PSS in 1-2 years, what is the primary care physician supposed to do with that information?

Thank you for your question. The third and fourth paragraph of the introduction (page 4) is directed at a rationale for why predicting PSS is important. In brief, detection of PSS is often delayed and this comes with a high burden on the patient and the health care system. Other screening tools seem to have been unable to improve prompt detection of PSS.

Since physician identification of PSS is often delayed, we applied a 'prediction gap' of two years. Although studies show that detection may be delayed longer (as seen in fibromyalgia; Gendelman et al., 2018) this may differ between PSS-subtypes (see also Comiskey & Larkan, 2010) and we assumed that the greater the prediction gap, the less likely a good predictive accuracy could be expected. Elevated risk of PSS may thus in actuality be detecting patients who already have PSS for some time.

To clarify our rationale, we have added the following at page 4: "Furthermore, research shows that a timely 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." (see for example Henningsen et al., 2018)

4. The main findings hinge on their definition of PSS, which is "based on previous research" but should be explained more fully in the present paper. The authors acknowledged that their PSS case ascertainment method was "suboptimal" because incidence rates were lower than expected, but I don't see reference 7-year incidence rates reported, only prevalence rates. To explain their lower than expected incidence rate, they offer "we cannot be certain that symptoms are not fully explained by a medical disorder." This could explain overestimating, not underestimating, incidence. We are still left without any possible explanations for the low rate of PSS.

Thank you for bringing up this point. For this study, we used conservative methods for selecting PSS cases to be as sure as possible that we selected cases and not symptoms that can be explained by a medical disorder. Because this has been a primary focus, we may have overemphasized this point. Furthermore, in the case of this study, a 2-year incidence rate is reported (cases are selected in 2017-2018), although this is not an accurate incidence rate, because we excluded a large part of the population.

In all, we have corrected our description as follows (page 12): "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, [53] and data-driven selection would enhance re-usability of routine care data."

5. It should also be noted that the low incidence of PSS is even more surprising given that only patients with at least 10 prior clinic contacts were eligible for this study.

Thank you for your comment. This is indeed the case; as mentioned in the discussion and explained in the previous point, the selection method is suboptimal (page 12). We aimed to get a varying group of patients with PSS (i.e., the broad spectrum, independent of getting a syndrome classification, an ICPC-code, etc.), but did not want to introduce too much error by including patients with persistent symptoms that are primarily explained by a well-understood biomedical disorder. Therefore, the selected group was rather too specific then too sensitive. This may also partially explain why the predictive accuracy is limited.

6. More than half of patients with PSS had a physical comorbidity. How can the authors be sure that the functional somatic symptoms were not fully accounted for by the physical comorbidities? Again, this gets at the PSS case ascertainment method. If the primary outcome is not truly PSS, that would

explain some of the surprising null findings (number of clinic visits not associated with PSS). The authors should consider re-running their prognostic models in just the patients without physical comorbidities.

> Thank you for this comment. Historically, PSS was diagnosed by exclusion of physical conditions. Recent developments in the field indicate that patients with physical comorbidity can also have PSS (this is quite common, as seen in this and other studies). Excluding patients with physical comorbidity would therefore exclude a big part of the population and goes against the current trend in the field (see Löwe et al., 2021).
>
> Regarding the null findings for number of consultations, the change in the predictive value may explain this because adding latent variables that explain the relation between PSS and consultation frequency decreased the predictive value of consultation frequency. This was previously not well described. The following sentences have been adjusted and added to describe this in the results: "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." (page 10); and discussion: "…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." (page 13).

7. Patients who already had PSS at the beginning of the observation were excluded. Wouldn't this bias the sample to younger patients and those with relatively mild, transitory PSS?

> Thank you for this question. We don't think this is a problem because PSS is generally 'diagnosed'/registered with a large delay. This would mean that most patients already have PSS for several years when they are identified/registered by their GP. Although we may have a relatively young sample, compared to the population of patients with PSS, the aim of this study is to identify patients at risk of PSS onset. To clarify we have made a small change to the description in the methods section (page 6): "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."

Minor concerns:

8. -Please state rationale for why these cut-offs were chosen (1) registered at least 7 years at the same general practice, (2) >10 clinic visits, and (3) 2 year prediction interval

> Thank you for your comments.
>
> (1) We deemed the use of 5 years of data for detecting predictors desirable and needed at least two years for the prediction gap. Therefore, we needed 7 years of data. If patients were enrolled for a shorter period, we have a higher chance of missing information.
>
> (2) The use of routinely collected clinical data for research purposes comes with many challenges. One of these is that patients need to visit their GP for us to have data to analyse. While we are unaware of a rule of thumb, we were sure that having less than 10 contacts over a 7-year period would provide too little data to analyse.
>
> (3) Please see response to comment 3.

See page 6 under Study design and a small textual adjustment under Study population ("These criteria were used to ensure availability of enough registrations per patient to enable candidate predictor construction.").

9. -"While most of these symptoms are self-limiting,…" (page 4, line 13). What follows does not seem connected to this opening statement.

Thank you for helping us clarify. What we meant to say was that since most of the symptoms are self-limiting (i.e., symptoms pass by themselves), they are not a problem, but identifying what symptoms will become problematic is more challenging. To clarify, we have adjusted the sentence as follows: "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.(Murray et al., 2016)"

10. -"Lastly, some physicians refrain from using terms beyond the biomedical domain for somatic symptoms" (page 4, line 27). Not sure what this means.

Thank you again for helping us clarify. We have adjusted the sentence as follows: "Lastly, some physicians refrain from using terms beyond well-established biomedical disorders for somatic symptoms"

11. -46.8% of the group without PSS had a mental health disorder. This seems very high. Was this an expected finding, consistent with known epidemiology?

Thank you for your comment. Indeed, this prevalence rate is higher than epidemiological findings (for example, Baumeister & Härter, 2007). Our definition of the "mental comorbidity" category include ICPC codes in the P-chapter (psychology) and ATC-codes for antidepressants and antianxiety medications. The P-chapter includes both symptoms and disorders, but these may be hard to differentiate, since GPs may be inconsistent in the use of these (e.g., using symptom codes in case of a disorder and vice versa). Therefore, we decided to include all the P-codes. We agree that our descriptor of the category is incorrect and have adjusted the name of the category in the manuscript to "mental health complaints".

12. -The finding that stable lab test results were associated with onset of PSS is interesting and could be elaborate on

We agree that this is an important finding (that also emphasizes the validity of our outcome), especially since adding the additional bootstrap analysis (see comment 14). We therefore added the following sentence (and in accordance with the findings from the additional bootstrap analysis replaced) (page 12): "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."

13. -The Patient and public involvement statement requires elaboration. Who did the authors consult with, when, and about what. What was the outcome of the consultations?

Thank you for this suggestion. We have adjusted this section as follows on page 8: "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 were 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."

14. -Please add 95% confidence intervals to Table 3

We appreciate this comment, which is a common request. However, the use of 95%CI's is not deemed reliable and unbiased in regression analysis which include variable reduction. Since predictors are already preselected, the variance is changed and the predictor estimates are biased (in the case on LASSO regression, estimates are reduced, which also explains the low ORs). In essence, the presentation of predictors is secondary to the aim of the analysis; this analysis is directed at finding the best possible model to predict PSS onset (i.e., predicting vs. explaining, see Shmueli, 2010). However, since it is clinically interesting to see what predictors comprise the best model, we present them in table 3. We consulted a statistician because we agree that the presentation of the predictors should include some verification that these are truly important. The statistician proposed we evaluate the stability of the results by running the analysis on bootstrap samples. Therefore, we have run the analysis 1000x over bootstrap samples and have included a column that presents the percentage of times a predictor was not reduced to zero by the LASSO regression. The percentage added to table 3 represents the stability of the predictors for predictive modeling of PSS onset based on routine care data. Percentages indicate which predictors are essential and which are exchangeable. For this addition, the following has been added to the methods section (page 8): "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." The following has been added to the text of the results section (page 10): "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)."

Reviewer: 2
Competing interests of Reviewer: None

****

Reviewer: 3

Dr. Sebastian Kohlmann, University Medical Center Hamburg-Eppendorf

Comments to the Author:

Short summary:

The manuscript reports on a novel and interesting approach to early identify patients with persistent somatic symptoms (PSS) in primary care. The registry-based dataset is large and the analysis is complex. The results indicate that GP-based data can moderately contribute to predict the onset of

PSS. The authors conclude that clinical decision rule based on structured symptom/disease- or medication codes can efficiently identify patients with PSS.

General comment:
The manuscript is an important contribution to the field of PSS and clinically relevant. In most parts it is well-written (introduction and discussion); in other parts the manuscript would benefit from presenting more information and illustration of the analysis (methods, results). The analysis appears to be sound but should be reviewed by expert in the field of machine learning. Strengths of the manuscript include its large dataset, the novel approach to identify PSS and the complex analysis. The main limitation, in my view, is that no patient-reported data was used in the prediction models. Additionally, I would slightly disagree with the authors' conclusion that routine primary care can be used to early predict PSS. The AUC values of the presented models can be judged as moderate (only). The results indicate that the early detection of individuals can be enhanced with data from primary but still there is large room for improvement. Below, I have made some suggestions to improve the manuscript.

Major strengths:
Innovation and relevant research idea
Large dataset

Major weaknesses:
No patient reported outcomes in the data set

Abstract:
15. The readability of the abstract needs to be improved. The information is very hard to follow: there are many abbreviations and the analysis and the dataset is complex. The conclusion is not correct and should be formulated more cautiously.

> Thank you for this suggestion. We have adjusted the abstract to make it more understandable. Due to the journal guidelines, it was difficult to explain everything properly. Adjustment of headings was needed to explain the methods more thoroughly.

Introduction:
16. The introduction is well written. The following papers may add some knowledge to recent developments in screening questionnaires and the current health care situation:
Toussaint, A., Hüsing, P., Kohlmann, S., & Löwe, B. (2020). 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). Psychological Medicine, 50(2), 324-333. doi:10.1017/S003329171900014X
Kohlmann, S., Löwe, B., & Shedden-Mora, M. C. (2018). Health care for persistent somatic symptoms across Europe: a qualitative evaluation of the EURONET-SOMA expert discussion. Frontiers in psychiatry, 9, 646.

> Thank you for bringing our attention to these valuable articles. They have been added as references on page 4. Including an additional sentence: "Research from a European network of experts in the field stresses the need for a systemic change to overcome these challenges.[33]"

Methods:
17. Why was having 10 or less contacts with general practice an exclusion criterion?

Thank you for giving us an opportunity to clarify. The use or routinely collected clinical data for research purposes bring with it many challenges. One of these is that patients need to visit their GP for us to have data to analyse. While we are unaware of a rule of thumb, we were sure that having less than 10 contacts over a 7-year period would provide too little data to analyse.

18. With respect to the SPADE algorithm: Why was a “1% difference” chosen as relevant?

This was a data-driven decision. The 1% difference in support for a sequential pattern indicates that the pattern should at least be detected 1% more often in one of the cohorts. Due to the nature of routinely collected data, which has high levels of non-random missing values, frequencies of variables are much lower than we would have in data that has been actively collected for research purposes. Unsurprisingly, frequencies of sequential patterns that we found in the data were also low (=support level). For example, the highest support level for the multi-sequence patterns is ~10% and the highest differences between the PSS- and non-PSS-cohort were 2.2%. The cut-off of 1% allowed us to include 20 multi-sequence patterns (and an additional 37 single-sequence variables). Given these low difference scores it may not be surprising that all estimates of multi-sequence pattern were reduced zero in the LASSO regression. Ultimately, we could write a full paper on this part of the study but given the limited contribution to the final model we decided, for the purpose of this paper, to limit the focus on this method. To help clarify this complex method for our reader we have added the following in the methods section (page 7): “…the support value (i.e., prevalence of the pattern in de dataset). Please see (Zaki, 2001) for a more detailed description of the SPADE algorithm.”.

19. Appendix S2 is not referenced in the text. Title of appendix 2 should be more precise: What is statistically meant with “most important patterns”?

Thank you for bringing this to our attention. The reference was included on page 9, but not very clearly. In data science and machine learning, instead of speaking of statistically significant predictors or ORs, outcomes are generally presented as ‘most important predictors’ and this is also how R presents the results by default, although it is not uncommon to get ORs as well. Since our target audience is medical professionals and health scientists, we decided to focus on giving more common parameters (i.e., ORs). While checking this point, we also adjusted the order of the tables to fit their mention in the manuscript. We have adjusted the title of Table S3 (formerly S2) of the appendix to “Patterns derived from the SPADE algorithm and subsequent LASSO regression for the sequential patterns model”.

20. A figure that illustrates the regression model(s) with outcomes and statistical assumptions could improve understandability of the method section.

Thank you for helping us clarify. Firstly, we have adjusted figure 1 to make the methods used clearer. However, we have not added the assumptions in this figure. The methodological differences between predictive research vs. explanatory research make it difficult to translate these analyses for clinical application. While most researchers in the field are more used to explanatory methods, for which traditional assumptions apply, in predictive research (such as when using machine learning techniques as done in our study) these assumptions are less important. When using machine learning techniques, we are not primarily interested in

explaining, but rather finding the best way to predict (as was the primary aim of this study). In this case, assumptions such as normality or multicollinearity are less problematic (although we did correct for multicollinearity to ensure we can present valuable predictors). In machine learning, overfitting of the training data is the main concern. (See Shmueli, 2010 for more details) Second, we have added a section on “final model evaluation” to clarify the method of validation further (page 8):”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.”

Results:

21. The sample description should include statistics on the differences between patients with and without PSS.

Thank you for this suggestion. Assuming you mean that the characteristics of the non-PSS cohort is missing and that p-values need to be added. We have consulted a statistician about this point and he confirmed that reporting of the total cohort is more common as this represents the general population. Furthermore, given the large sample size and skewed distribution of patients with (1%) and without (99%) PSS, the characteristics of the total cohort will be similar to that of the patients without PSS. Due to the large sample size, the p-value statistics on the differences between patients with and without PSS have limited value (i.e., any difference will be significant). At present we are unaware of a good alternative (although Gomez-de-Mariscal et al., 2021 may have a solution in the future) and therefore decided to refrain from reporting a statistic.

22. Positive and negative predictive values should be mentioned in the text and table 2.

Thank you for this suggestion. In table 2 we presented the sensitivity and specificity of the model. We believe this is a better and more commonly used way to present the patient classification, especially because it is independent of prevalence.

23. Confidence intervals should be added with respect to the final predictor set (text and table 3).

Please see comment at point 14.

24. Are there any indicators to judge the validity of the model? The results would benefit from presenting a sensitivity analysis. E.g., What would the models’ properties be like if the outcomes were analyzed separately (ICPC-codes for PSS-syndromes vs. PSS-umbrella terms vs. 4DSQ? Another approach to show the validity of this diagnostic approach would be to show that the model’s prediction is comparable to established risk models to predict the course of somatic diseases, e.g. the “EURO Score” for coronary heart disease.

Thank you for your suggestions. To evaluate the value of the predictors presented in table 3, we have performed bootstrapping, as described at comment 14. This enhances the knowledge on the validity of predictors.

We have also considered your suggestions. The difficulty with PSS is that its definition is ambiguous and gold standard diagnostics are unavailable. While we initially aimed to validate the patient group by sampling patients and asking their GP to diagnose the patient, research (for example Warren & Clauw, 2012) and preliminary work from our group (Kitselaar et al., 2021a) showed that GPs find it difficult to identify these patients. Kitselaar et al., 2021a also

shows that GPs differ greatly in their methods of PSS registration. Other studies also show that GP registration is often scattered (for example Pohontsch et al., 2018). Therefore, we think doing a sensitivity analysis on the different registration based-PSS-subgroups as suggested does not seem viable. I.e., given the variation in registration behaviour, could we not expect GPs that register PSS a certain way also register other complaints/etc. a specific way? Since this seems likely, or at least not unthinkable, it should be expected that the prediction models for each PSS selection method would look different. We believe a strength of this study is that we have combined these methods and therefore filtered out some of this registration-based bias and variation. The findings in Kitselaar et al., 2021b show that there is a good likelihood that we are identifying a patient group with similar and persistent problems. We think this is especially apparent in “Method B”(which we did not include in our selection methods for this studies outcome), where you see that patients have markedly higher rates of physical comorbidity and consultation frequencies (and other health care utilization variables). We therefore concluded that this method contains too much error and are more likely to contain many patients with physical disorders explaining their health care utilization. In sum, we think, with the current data further judgements of validity are not useful, but we are confident we are identifying patients who have an abnormally high level of persistent symptoms that would require a different treatment approach.

Discussion:
25. The discussion section is well-written and presents strengths and limitations. With respect to the main results, the authors’ interpretation should be done more cautiously: “Our study shows how routine primary care data can be used as a source that enables early prediction of PSS.” The AUC of the best model can be judged as moderate; specificity and sensitivity are not optimal to identify and rule out cases. Still, the results are important as this machine learning approach could be enriched with data from patients (PROMS) which might lead to better predictions. The authors should elaborate on the limitation that no PROMS were used and give further outlook with respect to existing studies on the criterion validity of available screeners (SSS-8, SSD-12, PHQ-15, DSQ-4, BDS, etc.).

Thank you for this evaluation and the suggestions.

First, we agree that the interpretation should be more cautious and have adjusted the manuscript as follows: “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.”

Second, we agree that the use of PROMS would be a valuable next step. The present study was designed to study the possibility to make a fully data-based prediction model, without the need for additional data collection so that it could be useful for routine primary care data as is. However, we have adjusted the last paragraph of the manuscript (page 13) as follows to give more appropriate suggestions for future research: “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.”

Reviewer: 3
Competing interests of Reviewer: None

****

Baumeister H, Härter M. Prevalence of mental disorders based on general population surveys. Soc Psychiatry Psychiatr Epidemiol. 2007 Jul;42(7):537-46. doi: 10.1007/s00127-007-0204-1. Epub 2007 May 21. PMID: 17516013.

Comiskey, C., & Larkan, F. (2010). A national cross-sectional survey of diagnosed sufferers of myalgic encephalomyelitis/chronic fatigue syndrome: pathways to diagnosis, changes in quality of life and service priorities. *Irish Journal of Medical Science*, *179*(4), 501–505. https://doi.org/10.1007/S11845-010-0585-0

Gendelman, O., Amital, H., Bar-On, Y., Ben-Ami Shor, D., Amital, D., Tiosano, S., Shalev, V., Chodick, G., & Weitzman, D. (2018). Time to diagnosis of fibromyalgia and factors associated with delayed diagnosis in primary care. *Best Practice & Research Clinical Rheumatology*, *32*(4), 489–499. https://doi.org/10.1016/J.BERH.2019.01.019

Gómez-de-Mariscal, E., Guerrero, V., Sneider, A. *et al.* Use of the *p*-values as a size-dependent function to address practical differences when analyzing large datasets. *Sci Rep* 11, 20942 (2021). https://doi.org/10.1038/s41598-021-00199-5

Henningsen, P., Zipfel, S., Sattel, H., & Creed, F. (2018). Management of Functional Somatic Syndromes and Bodily Distress. *Psychotherapy and Psychosomatics*, *87*(1), 12–31. https://doi.org/10.1159/000484413

(a) Kitselaar, W. M., van der Vaart, R., van Tilborg-den Boeft, M., Vos, H. M. M., Numans, M. E., & Evers, A. W. M. (2021). The general practitioners perspective regarding registration of persistent somatic symptoms in primary care: a survey. *BMC Family Practice*, *22*(1). https://doi.org/10.1186/s12875-021-01525-6

(b) Kitselaar, W. M., Numans, M. E., Sutch, S. P., Faiq, A., Evers, A. W., & van der Vaart, R. (2021). Identifying persistent somatic symptoms in electronic health records: exploring multiple theory-driven methods of identification. *BMJ Open*, *11*(9), e049907. https://doi.org/10.1136/bmjopen-2021-049907

Löwe, B., Levenson, J., Depping, M., Hüsing, P., Kohlmann, S., Lehmann, M., Shedden-Mora, M., Toussaint, A., Uhlenbusch, N., & Weigel, A. (2021). Somatic symptom disorder: a scoping review on the empirical evidence of a new diagnosis. *Psychological Medicine*, *52*(4), 632–648. https://doi.org/10.1017/S0033291721004177

Pohontsch, N. J., Zimmermann, T., Jonas, C., Lehmann, M., Lowe, B., & Scherer, M. (2018). Coding of medically unexplained symptoms and somatoform disorders by general practitioners - an

exploratory focus group study. *BMC Fam Pract*, *19*(1), 129. https://doi.org/10.1186/s12875-018-0812-8

Shmueli, G. (2010). To explain or to predict?. *Statistical science*, *25*(3), 289-310.

Warren JW, Clauw DJ. Functional somatic syndromes: Sensitivities and specificities of self-reports of physician diagnosis. *Psychosom Med* 2012;74:891–5. doi:10.1097/PSY.0b013e31827264aa

Zaki, M.J. SPADE: An Efficient Algorithm for Mining Frequent Sequences. *Machine Learning* 42, 31–60 (2001). https://doi.org/10.1023/A:1007652502315

VERSION 2 – REVIEW

REVIEWER	Silverberg, Noah The University of British Columbia
REVIEW RETURNED	11-Dec-2022

GENERAL COMMENTS	The authors addressed all of my concerns.

REVIEWER	Kohlmann, Sebastian University Medical Center Hamburg-Eppendorf, Department of Psychosomatic Medicine and Psychotherapy
REVIEW RETURNED	05-Dec-2022

GENERAL COMMENTS	The authors thoroughly addressed most of my comments. Still, the following points should be addressed: Abstract: The conclusion is still not correct. It should clearly state the finding: there is a 70% chance that the model will be able to distinguish between individuals with / without PSS. This can be interpreted as low to moderate diagnostic accuracy. The paper did not investigate, whether this predication model is an efficient way to support GPs in identifying cases. Results: With respect to the sample description, the author report differences (e.g. "Compared to the total cohort, patients with PSS are more likely to be female (69.0% vs. 52.9%)"). But no p-values are reported. If the authors state that there is a difference, they should indicate that with p-value. The authors argue that their prediction model is relevant for clinical practice. Thus, positive and negative predictive values should be mentioned as these reflect the performance in the population accounting for the prevalence.

VERSION 2 – AUTHOR RESPONSE

Dear reviewers,
Thank you for the time and effort put into reviewing our manuscript. We have considered the final comments and we have adjusted our manuscript accordingly. Please find a detailed description of the adjustments below.
On behalf of all authors.

Reviewer: 3
Dr. Sebastian Kohlmann, University Medical Center Hamburg-Eppendorf
Comments to the Author:
The authors thoroughly addressed most of my comments. Still, the following points should be addressed:

Abstract:
The conclusion is still not correct. It should clearly state the finding: there is a 70% chance that the model will be able to distinguish between individuals with / without PSS. This can be interpreted as low to moderate diagnostic accuracy. The paper did not investigate, whether this predication model is an efficient way to support GPs in identifying cases.
Thank you for the suggestion. The conclusion has been nuanced accordingly (p.2):
“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.”

Results:
With respect to the sample description, the author report differences (e.g. "Compared to the total cohort, patients with PSS are more likely to be female (69.0% vs. 52.9%)"). But no p-values are reported. If the authors state that there is a difference, they should indicate that with p-value.
Thank you for your comment. We previously did not add the p-values because these have low value in large datasets. However, we agree that some statistic is required, and a good alternative is currently not available, so we decided to add them (page 9).

The authors argue that their prediction model is relevant for clinical practice. Thus, positive and negative predictive values should be mentioned as these reflect the performance in the population accounting for the prevalence.
Thank you for this suggestion. The positive and negative predictive values have been added on page 9 (“PPV is low (ranging from 1.5% to 1.7%) and NPV is high (ranging 99.5% to 99.6%).”).

We think this is a valuable contribution, since it corroborates one of our main discussion point so we made a small adjustment in the discussion too (p. 13): “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.”

Although we agree that this an informative addition and it is relevant information for clinicians, there is some debate on the accuracy of the measures. Therefore, we’ve added the following to the methods section (p. 8): “To evaluate the predictive value of each model, a sensitivity analysis was performed. This included prevalence independent measures (i.e., sensitivity and specificity) and prevalence dependent measures (i.e., 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 53,54]”

Reviewer: 2
Dr. Noah Silverberg, The University of British Columbia
Comments to the Author:
The authors addressed all of my concerns.

Reviewer: 3
Competing interests of Reviewer: none

Reviewer: 2
Competing interests of Reviewer: None

References
53 Molinaro AM. Diagnostic tests: how to estimate the positive predictive value. Neurooncol Pract 2015;2:162–6. doi:10.1093/NOP/NPV030
54 Schelde AB, Kornholt J. Validation studies in epidemiologic research: estimation of the positive predictive value. J Clin Epidemiol 2021;137:262–4. doi:10.1016/j.jclinepi.2021.05.009