Peer Review File

Article information: https://dx.doi.org/10.21037/jmai-22-42

Reviewer Comments	**Response**
☒ I would recommend graphical improvements, especially fig. 6 needs to be more easily assessed.	We thank the reviewer for this comment and agree the figure could be much simpler. We have removed extraneous and information and streamlined the figure into a single dendrogram-heatmap with the experiment colorbars (k= 3 and 5) on the left and right of the y-axis respectively. We hope this is a clearer representation of an example of the results and how they are useful. Additionally, we found some misalignments in figure 3 which were in the shared pdf but not in the original figure – we will reach out to the editors to seek for advice.
☒ I personally prefer to read separately discussion and conclusions. The final part needs some shortening or split (discus+concl)	Thank you for your comment, we split the discussion and conclusion and integrated additional elements as suggest by reviewer 2.
☒ p1:14-18 EHRs are quite new and contain very heterogeneous data. Hence, computational methods have not consolidated in any way and it will take quite a while before methods become more harmonized and interpretable.	We agree with this assessment. This is also one of the motivations for this work and we hope that we can support this *consolidation* process over the next years. We extend the abstract: "…outcomes and remaining medical need. However, working with large EHRs dataset is still relatively new and contains various challenges due to the heterogeneous nature of the data. The recent interest of using ML-based aggregation of EHR data is mostly tool-driven, i.e., building on available or newly developed methods."

☒ p1:23 The dataset contained ...patients?	Typo Resolved
☒ p2:65 The difference between tool and method is very unclear. Please define these terms.	Thank you for pointing this out. We have used the terms “tool” and “method” interchangeably in the manuscript. We agree that this can cause confusion. We updated the manuscript and use now “method” consistently.
☒ fig.2 Stage 5 could introduce bias	We agree that this process can also introduce additional bias. Based on our internal experience, this step requires good cooperation between the data and clinical experts. We extended the manuscript in the method section to point this out by: “… This step relies on close interaction between clinical and data experts as e.g. the removal or modification of features might introduce new bias.”
☒ p6:104 The term "experiments" seems odd. Please specify what you mean in this context.	We added a definition of the term *experiment* within the context of the paper. We agree with your comment in general and had also internal discussions about that. However, we kept the term in absence of a better alternative. **“…**Note, within the context of this publication, an experiment is defined as applying a specific stratification algorithm on data of a patient cohort including specific preprocessing steps and algorithms parameters which results in a mapping of patients to different clusters.” (page 2)
☒ p8:157 It should be clearly noted that k=12 may not be sufficient for large patient cohorts with various expected subtypes. How do you actually select the optimal k? This	Thank you for pointing out the interesting paper of Rose et al., which we added as reference in our manuscript (in particular the error model is an interesting extension).

step could be automated with a more sophisticated method as in [].	We specifically selected the maximum number of clusters k_max = 12 to ensure that the identified clusters have sufficient patient numbers to be practically relevant for e.g. developing a specific therapy and focused on first 2 local maxima to balance number of relevant results and number of results which require clinical evaluation. This might be different for other applications. Plots of the silhouette score can be added if required, but we see limited values as they very across datasets and cohorts. Even though we agree, that identifying the optimal number of k is important, we want to point out that the approach presented by Rose et al. also has parameters to be specified such as the row_threshold and col_threshold which impacts the number of identified clusters and needs to be adjusted for the specific application. We added: "Note, the parameter depends on the dataset and specific application and are considered as an example. Here, we focused on smaller numbers of *k* to ensure that identified sub phenotypes have sufficient number of patients to be practical relevant and focus only on the best 2 results to effectively reduce results which require clinical evaluation. An alternative approach for this step was presented in (https://doi.org/10.1073/pnas.2118210119), which has an automatic determination of optimal number of clusters included."
☒ p9:159 Why do you only select	See answer above.

the first two local maxima? What is the coefficient "landscape" here? This, together with the selection of k number of clusters, might work for the shown sample data but seems to be rather weak from a data science perspective requiring generality and dataset independence.	
☒ p9 So, clusters are computed separately for the base and the surrogate model. Then you determine the average? This doesn't seem to be very robust for scenarios when base and surrogate model are very different, especially since the average may still seem to be alright. How do you deal with these scenarios?	We thank the reviewer for pointing this out. Finding the average is a simple example of how to interpret the results, however, we agree that this is not optimal and might cause more confusion in this example. We have therefore removed this. Main objective is to identify results which differ with respect to the outcomes (e.g. increased risk of mortality) using criteria which are simple to apply (e.g. in a clinical trial scenario). For this purpose, computing only the base and surrogate ranking scores and comparing the difference is required. We updated the “pseudo code” in the method section and adjust the results and Figure 7.
☒ p10:203 "Through"	Typo Resolved
☒ p11:220-222 Not sure what the consequence should be in this scenario. Is this cluster ignored for further interpretation or is this cluster the most interesting for further investigation?	From our point of view, if a potentially interesting clusters is defined based on non-meaningful features, the specific features should be either removed or combined with other features to make the feature clinically meaningful (see feature curation). One example is shown in figure 11, where in the first instance procedure code of “radiology of one body area” was identified as most relevant feature, which is very

	unspecific. Consequently, it was removed from the analysis (figure 12). Of course the removal features and repetition of the analysis might result in different cluster results.
☒ p16:325-329 The description of the filtering criteria is done very well. However, as a result, what had been over 800,000 patients in the beginning are now a little over 4000 (0.5%). Still a large number, but not quite having the same statistical significance. I think it should be made more clear how many patients were actually used as an input after filtering.	Thank you for pointing this out, we added additional references to Table 1, where all 4 cohorts are summarized, in the "clustering methods" section and at the beginning of the result section. e.g.: "…we illustrate with examples on a clinical study using a large-scale EHR dataset (using the cohorts defined in 錯誤！找不到參照來源。) which focuses…" Further, the patient numbers are mentioned in the abstract.
☒ p16:336 "that a relevant"	Typo Resolved
☒ p17:349 I guess the age of death should be greater than the age at first admission.	We corrected the mistake.
☒ p18:365 I would not refer to this as an embedding space. That said, the transformation methods are explained afterwards, so it is quite clear what is happening.	We have modified the text and replaced embedding with data representation: e.g.: "Before clustering, the preprocessed data is transformed into one of the following three data representations…"
☒ p19:387 The formula for the Jaccard index is rather unintuitive: sum_1tok sum_1tol nij should be equal to 1, hence it is simply overweighting the intersection	We thank the reviewers for bringing this to our attention. We have modified the equation to a more standard format.
☒ Methods section: The methods section is quite comprehensive but I am not sure I could actually reproduce the whole procedure	From a high-level perspective, we are describing a novel workflow / framework consisting of established methodologies. The novel aspects we

based on the given information. It is also not entirely clear to me what the authors' own contribution really is and what they have used from literature. Please make sure this is clear and that all used methods are referenced if not already done so.	describe in this paper is how to conceptually apply these to facilitate greater engagement of non-data scientists (e.g., typical clinicians), and to enable greater throughput of analyses by reducing clinical input to key steps where we also provide guidance on how to assess relevance (e.g. see our feedback here and revisions of the manuscript for the surrogate models). The described methods should be considered as examples within this new framework and are generally available (e.g. surrogate models). For meta-clustering and pattern screening, we extended the manuscript with pseudocode, as we consider these as "non-standard" methods. We have extended the Method section to now include information about software frameworks or packages.
☒ p22:433 "setS"	Typo Resolved
☒ p23,p24 results The presented results are quite surprising to me since they do not seem to serve as a convincing example: 1. With regard to the description, doesn't this mean that the interpretation of the clusters is quite impossible? 2. With regard to the description, doesn't this mean that the surrogate model exhibits problems with such complex data? 3. A decrease in the bleeding score and an increase in mortality seems counterintuitive to me. 4. It is stated that the combination of base and surrogate model is causing e26c3 to be captured in the pattern screen. Isn't this a good example for	Thank you for pointing this out and we made major adjustments in the manuscript (see also the comments regarding "averaging" on p9). The main objective of the pattern screening approach is to support clinicians to quickly identify which clusters are potentially relevant. To emphasize this, we added (amongst others): "In contrast of analysing all 71 cluster results individually, **錯誤! 找不到參照來源。** provides a quick overview which cluster is relevant with respect to the outcomes. ..."

my concerns related with the described scenario about the averaging of results mentioned above (p9)? And is the surrogate predicting the same clustering? I need more details about base model predictions and those of the surrogate model to fully understand what is happening.	The pattern screening approach is very flexible in principle and depends strongly on the clinical questions / objective of the study. To make the clinical questions of our example case study more explicit we added in the "use-case method" section: "Clinical Questions The objectives of this patient stratification study are: I. Can we identify clinical meaningful sub-phenotypes of patients within patients who have a first diagnosis of ischemic stroke or an acute heart failure episode? II. Do these sub-phenotypes differ with respect to their clinical outcomes such as mortality rates? III. Are these sub-phenotypes practically relevant, meaning that they can be defined by using inclusion and exclusion criteria with a high degree of clinical meaning to define the population? The implemented pattern score aims to address the 2nd and 3rd clinical question. Which cluster has the biggest difference with respect to one of the clinical outcomes and can this cluster be implemented using a simple surrogate model (comparison between base and surrogate model) To answer your questions: 1. We made major adjustments to the results and discussion section. As said above the objective is not to interpretation

	of the clusters, rather pinpointing which cluster could be most relevant. 2. Yes, we experienced in many cases that the surrogate model was not capable of capturing all interactions between the different data items. However, we want to emphasize that the use of the surrogate model was driven by the third clinical questions, the requirement that a novel patient cluster can be defined by simple inclusion / exclusion criteria as in clinical trials. Therefore, we did not implement further complex surrogate model beyond a decision tree. We addressed this in a new section in the discussion section: "Which specific method is used as surrogate model, depends strongly on the objectives of the study. We have used a simple decision tree, which matches the requirements of the third clinical question, meaning that a relevant cluster can be defined by a few clinical meaningful inclusion and exclusion criteria. This is particularly useful in the context of clinical trials, where each additional patient selection criteria can have a big impact on patient recruitment. However, there are several drawbacks in using tree-based methods, chiefly among them are the inability to handle temporal data. In our example, all temporal feature data was aggregated before it was applied to the surrogate model. In future work, there is scope to develop surrogate models that can accommodate patient trajectories. This

	resulted partly in low accuracy values of the trained surrogate models and big difference between the pattern screening scores of the base and surrogate model for some cluster results (see **錯誤! 找不到參照來源。**, exp 26 k=3). However, if the study objective focuses, for instance, only on understanding which clinical parameters are relevant, there are several well-developed more advanced surrogate models , such as Ripper(41), Trepan(42), or RuleFit(43) – the details of which are beyond the scope of this study - which could result in an better overlap between base and surrogate model." 3. Yes, we agree. The surrogate model was not able to capture the characteristics of the original cluster. Due to the big difference between the base and surrogate model, this cluster was not further investigated. 4. As mentioned above, we removed the computation of the average within the manuscript as it is not relevant for the analysis and created unnecessary confusion. To evaluate the performance of the surrogate model, there are from our perspective two relevant metrics: a) evaluation of the accuracy metric (as mentioned in the surrogate section - describing how good the surrogate could predict the same patients as in the base cluster) and b) comparing pa ttern screening values between base and surrogate model (see

	Figure 7 – which indicates if the surrogate cluster has maintained similar outcome characteristics).
☒ p26:501 "TO predict"	Typo Resolved
☒ Results: I am not completely convinced that the decision tree surrogates are actually working sufficiently in general. It is even stated that "the surrogate model's predictive performance was variable in practice, however in some cases the surrogate model was able predict the original clusters with a high degree of accuracy, even at a very low tree depth."	Please see the response for comment p32:579 below.
☒ p27:521 I think the term "unknown criteria" is not quite correct since this would describe a novel discovery. Rather, it revealed "not previously thought of criteria". I guess it becomes clear to the reader what the authors mean, but I suggest to be more precise.	We agree with this comment and changed the text to “previously not considered criteria”.
☒ p32:578 The term "tools" seems odd and not in line with previous usage. Please define in the beginning what you really mean. Here it seems like "strategy" would be a much better term. Going back to the publication https://doi.org/10.1073/pnas.2118210119 the authors provide a web-service and a stand alone version, so they provide "tools" that utilize their "method".	We changed the term to “*strategy*” and as stated above, we replaced the term “*tool*” by “*method*” in the whole manuscript.
☒ p32:579 Based on the results the authors present I cannot see that they "show the power of surrogate	We have amended the methods, results and discussion sections as mentioned in the comments above. In the specific

modelling for explaining patient phenotyping". This seems to be quite overstated.	case, we adjusted the text to “we show how surrogate models can be used to explain patient phenotypes”. As mentioned in the comments above, the objective of the manuscript was not to evaluate different surrogate models, rather presenting an overall framework for such type of analyses.
☒ p32:583 I think this has not been shown adequately.	Our main intention is to emphasize that this framework works with all possible (also deep learning based) patient stratification algorithms. We reformulated the statement to: “Significantly, the proposed framework is independent of the used patient stratification methods and does not put any constraints on the complexity of the stratification methods.”
☐ p34:631 One of the identified issues in this review is the performance of the surrogate model. Here, the authors list several promising alternatives to the presented decision tree. I strongly suggest to include results using these mentioned approaches and others for the surrogate model. This would also greatly strengthen the impact of a possible future publication.	Our goal was not to benchmark different surrogate models against other data / machine learning models, but rather to introduce a novel way / framework to add clinical meaning to initial results in a way that allows for checking whether or not change in parameters (from original to surrogate) preserve differences in outcomes. The selected surrogate model was motivated by the third clinical question, which focuses on defining new clusters by a few clinical meaningful concepts. Additionally due to the simple interpretability, it was a useful tool to build trust in the initial patient stratification results and to evaluate and optimize (see feature curation) the selected clinical concepts.
☒ Discussion: The details that I have been missing throughout the results	As discussed above, we extended the method section to provide a

section are consequently also not discussed, e.g. a critical discussion of the chosen cluster size among other details (see above).	justification for the evaluated number of clusters in the analysis and added a reference to your proposed paper. Additionally, we extended the discussion by: "A further challenge of the proposed meta-clustering approach is the selection of the optimal number of meta-clusters. Further approaches to automate this should be investigated such as proposed in (25). "
☒ There is no tool and I don't know how anyone (even programmers) would be able to apply it. The Date Sharing Agreement says "TBD". Since this is a method paper, at least the source code needs to be fully accessible via a git repository or similar. A tool should be made available to allow other researchers to actually apply the presented method to custom data.	As mentioned above, we replaced the word "*tool*" with *"method"* as we don't want to create a misleading impression of the manuscript. The focus of the manuscript is to present a new framework with additional steps to support clinical evaluation of patient stratification results. We consider the presented methods as examples within this framework and hope to encourage the community to develop further methods. Therefore, we don't not create a git repository. However, we extended the manuscript with pseudocode for meta-clustering and pattern screening to facilitate implementation. We updated the data sharing agreement, thank you for pointing out this mistake.