CTU
CZECH TECHNICAL UNIVERSITY IN PRAGUE

THESIS REVIEWER'S REPORT

I. IDENTIFICATION DATA

Thesis title:	**Visualization and analysis of patients digital phenotypes**
Author's name:	**David Kolečkář**
Type of thesis :	master
Faculty/Institute:	Faculty of Electrical Engineering (FEE)
Department:	Department of Computer science
Thesis reviewer:	Michal Huptych
Reviewer's department:	Czech Institute of Informatics, Robotics and Cybernetics

II. EVALUATION OF INDIVIDUAL CRITERIA

Assignment — **ordinarily challenging**

How demanding was the assigned project?

Assignment corresponds to the master thesis prerequisites. The assignment reflects an interesting and actual topic, and I evaluate the difficulty of the assignment as ordinary.

Fulfilment of assignment — **fulfilled**

How well does the thesis fulfil the assigned task? Have the primary goals been achieved? Which assigned tasks have been incompletely covered, and which parts of the thesis are overextended? Justify your answer.

The student respected all points of the assignment and fulfilled the assignment's demands.

Methodology — **correct**

Comment on the correctness of the approach and/or the solution methods.

I have several comments on the methodology. It is important to understand data. However, I did not find out the importance of this analysis for some interpretation of the obtained phenotypes. In other words, it is correct that the author performed the univariate analysis, but its outputs are not further processed or used.

The multivariate analysis includes only two graphs for age vs education background by sex and cigarettes per day vs town size by sex. There is not any multivariable analysis, such as logistic regression, that could show the importance of particular parameters in the outcome.

The clustering methods are described appropriately, given the focus of the work. I have comment on the methodology of the results and their interpretations. In fact, the phenotypes are created based on k-means clustering. It is clear that the k-means had better results in the observed criteria. However, it would be great to see what it means in the phenotype structure, not only in the numbers of defined criteria. In my perspective, there is a lack of effect size and explainability of the results. The author just states two specific phenotypes (the best and the worst probability of therapy finishing) in the conclusion. However, the differences between the phenotypes are often relatively small and no explanation or justification is described for these results.

Technical level — **B - very good.**

Is the thesis technically sound? How well did the student employ expertise in the field of his/her field of study? Does the student explain clearly what he/she has done?

The technical level of the work is on high level, although I have a few comments on this:

In Table 13, it would be much better to include to the table also some representation of the particular features (e.g. median and IQR). Moreover, it is not good mention "mean" and define testing by Kruskal-Wallis test.

Further, I do not understand why author use mean and std to represent numerical values when above in the work he strongly rejects normality of the data. Moreover, in the phenotypes is not use any range for the numerical values (which can be related with my previous comment).

Formal and language level, scope of thesis — **A - excellent.**

Are formalisms and notations used properly? Is the thesis organized in a logical way? Is the thesis sufficiently extensive? Is the thesis well-presented? Is the language clear and understandable? Is the English satisfactory?

CTU
CZECH TECHNICAL UNIVERSITY IN PRAGUE

THESIS REVIEWER'S REPORT

I find the way of aligning the captions under the images strange.

Selection of sources, citation correctness **A - excellent.**

Does the thesis make adequate reference to earlier work on the topic? Was the selection of sources adequate? Is the student's original work clearly distinguished from earlier work in the field? Do the bibliographic citations meet the standards?

Selected sources are sufficient for the scope of the thesis, and the use of references is correct and appropriate.

Additional commentary and evaluation (optional)

Comment on the overall quality of the thesis, its novelty and its impact on the field, its strengths and weaknesses, the utility of the solution that is presented, the theoretical/formal level, the student's skillfulness, etc.

III. OVERALL EVALUATION, QUESTIONS FOR THE PRESENTATION AND DEFENSE OF THE THESIS, SUGGESTED GRADE

David Kolečkář prepared the analysis and pipeline to generate digital phenotypes for the successfulness of chronic nicotine addiction therapy. The work is based on data from smartphone application. He processed the overview of related methods including methods for embedding of nominal categorical values for using in the clustering methods. Student prepared and implemented pipeline for the embedding and clustering of parameters.

I have comments on the methodology based on meaning of the phenotype and definition of the effect size of the results and their better explanations. I have also a few smaller comments on the technical level. In summary, I evaluate the thesis as very good.

The grade that I award for the thesis is **B - very good.**

Questions:

1) How often are data incomplete, and should the analysis omit the record? Was it a serious problem in your case for analysis?
2) Is it some down limitation in sample size for utilization of the entity embeddings and how was the embeddings time-consuming in your case?
3) You have used two datasets and the best and worst results were reached for the second dataset B, which is created as a reduction of the first dataset. Is it possible to explain why the second dataset was more useful in the identification of these phenotypes?

Date: **14.6.2023** Signature: