No evidence for communication in CLIS?
A report about NIRS-based BCI communication in CLIS

xxx Name of Whistleblower

09.10.2017

Abstract

In 2014, Gallegos-Ayala et al. have published a first case report of established communication in a CLIS patient using a NIRS-based BCI. This was followed up in 2017 by Chaudhary et al. showing evidence of NIRS-based communication in 4 CLIS patients, as well as results from online feedback sessions and open question sessions. In this report, I argue that this line of research suffers from a number of mistakes and that the claim of established communication in CLIS does not hold. I concentrate on 3 main points that substantiate this hypothesis: 1) a trigger-mistake in the pre-2014 data. 2) a wrongly applied parameter optimization in the 2017 publication. 3) errors in the presentation of the results from the online feedback and open questions sessions. As I don't have all data and information available, there are still some holes in my argumentation, which I will point out. At the end, I recommend the next steps that should be taken to get a definitive answer regarding the questions how much the mentioned errors influence the results and if there is indeed no evidence for communication in CLIS patients.

1. Pre-2014 data and the trigger problem

In September 2013 I got some NIRS data from Bin Xia to look at and to check his results. They were recorded on 22.7.2013 and 23.7.2013 from subject VP105. Bin Xia obtained very good classification results on this data, which I could confirm with accuracies of 88.7% and 92.5%.
To mark the "yes" and "no" segments in the data, trigger 4700 and 100 were used. However, trigger 100 is also used by the NIRS-machine to do a baseline reset. This was later confirmed by Bin Xia who checked the manual. If a baseline-reset happens in the "yes"-data, but not in the "no"-data, the data is contaminated by this technical issue and the classifier is likely trained to detect whether there was a baseline-reset or not. Meaning that the good classification results are due to a technical issue and not a hemodynamic response to yes/no.
I recommended to Bin Xia to re-record the data. As I only checked data from two sessions, I don't know, which other data is affected by this issue. I think that experiments with healthy subjects were done in 2013 and the data from Gallegos-Ayala et al. (2014) was also recorded in 2013, so that it is possible that the error persists in these data and was the reason for the above-chance classification accuracy.
I recommend to check the data used in Gallegos-Ayala et al. (2014) if those triggers were used and if this baseline-reset happened. It is important to check the raw data for this, not some preprocessed data. If a baseline-reset happened, the data should not be used. While an offline baseline correction would lead to a removal of this problem on first sight, it is not clear what the NIRS-machine does on baseline-reset. It is possible that filter artifacts are affecting the data after baseline-reset. In this case, the effect cannot be corrected offline and the data needs to be re-recorded.
My e-mail to Bin Xia with a figure showing the baseline-reset in the data can be found in the appendix A.1.

2. Parameter optimization wrongly applied in Chaudhary et al. (2017)

For Chaudhary et al. (2017) the NIRS data was classified using a Support Vector Machine (SVM). Details about SVM classification are not presented in the publication, but a grid-search was used to find optimal hyperparameters for the SVM (see Appendix A.2 for script). However, parameter-optimization and evaluation of classification accuracy are not separated. Instead classification accuracy was estimated with many different parameter combinations and the maximum value was reported, which leads to a serious overestimation of classifier performance.
As 12 values for C were tested and 20 values for gamma (note that gamma is only used for RBF-kernel and has

no effect with linear one, therefore should not be optimized here) a total of 12*20=240 evaluations were performed and out of the 240 classification accuracies, the best value was chosen and reported.
The effect can be better visualized when assuming a completely random data without any effect and the classification performance following a normal distribution. This would lead to an average classification accuracy >60% which would be significantly above chance with $p<0.005$, while a proper estimation (with parameter optimization) should result in 50% accuracy which is not significant.

2.1 Reanalysis of the raw data

I performed a reanalysis of the raw data. I applied the Beer-Lambert transform to obtain Oxy/Deoxy-values and bandpass filtered the data from 0.01 Hz to 0.3 Hz. As the raw data contains data for all sensor/emitter combinations, this results in 8*8=64 channels with 2 values each (oxy/deoxy) resulting in a total of 128 values per timepoint.
Data was cut into trials according to trigger 4 (no) and 8 (yes) and a 15s segment after each trigger was used for analysis.

2.1.1 Statistical analysis of the raw data

To check for significant differences in the data, a Wilcoxon's ranksum test was performed for each channel and each sample.

After correcting for multiple comparisons, there are no significant differences ($p>0.05$).

2.1.2 Classification of raw data

To estimate classification accuracy on this data, a 5-fold crossvalidation was performed for each day and patient. A SVM with a linear kernel was used. When no parameter optimization was performed and a default value of C=1 was used, an average classification accuracy of 49.66% was obtained (detailed results in Appendix A.3).
When using a (properly implemented) gridsearch to optimize the hyperparameter C, an average classification accuracy of 50.05% was obtained (detailed results in Appendix A.4).
Out of interest, I also performed a classification with a wrongly implemented gridsearch (similar to what was done for the paper), in which 17 different values of C were tested and the highest accuracy reported. In this case, an average accuracy of 59.33% was obtained. Please note that only 17 values were tested, not 240 as in the paper. The larger the number of evaluations, the more likely it is to find a better value and therefore higher values can be reported. Nevertheless, this kind of evaluation is methodologically not correct.

2.1.3 Discussion of raw data analysis

If there would be a hemodynamic response to yes/no questions in the data, it would have shown in this analysis. Classification accuracies around chance level and the lack of statistically significant differences show that there is no effect.
The methods used here, may not be optimal, but it should be emphasized that if there were a solid effect in the data, the method used here would show this effect.
Nevertheless, one could argue that a different kind of statistical test should be used. Chaudhary et al. (2017) showed a significant difference, but from the paper it is unclear how exactly this test was computed.
Further, the raw data with all sensor/emitter combinations was used, there is a lot of useless information. This may influence statistics in a sense that correction for multiple comparisons is harder. While it is not ideal for classification, it should not negatively influence classifier performance if the classifier is properly regularized, which was ensured by parameter optimization.
As a last argument, Chaudhary et al. used the slope of the signal, not the raw data. Again, this may influence statistical analysis. For the classification, a properly regularized classifier would be able to extract the slope out of the raw signal, so that classification accuracy should not be affected by this (or only in a minor way).

2.2 Analysis of preprocessed data

As a second attempt to reanalyze the data, the preprocessed data should be used. While Ujwal Chaudhary made the preprocessed data (.mat files) they used for classification available to me, there seemed to be a problem with some of these files (e.g. ALS_BCI2014\Patient3\matfile\visit1\Day1b1.mat), which is why this data was not used. Instead, the preprocessed data that was made available as supplementary files to Chaudhary et al. (2017) and published on Zenodo were used.
This data contained mat-files which contained two variables: `HbXtrueNIRSdata` and `HbXfalseNIRSdata`, which contained the preprocessed data to yes and no questions separated for each session.

2.2.1 Statistical analysis of preprocessed data

To check for significant differences in the data, a Wilcoxon's ranksum test was performed for each channel and each sample. After correcting for multiple comparisons (Bonferroni), there are no significant differences ($p>0.05$). As the preprocessed data consists of 20 channels with 93 timepoints each, the corrected significance level for 5% would be 0.05/(20*93)= 0.000027. Without correction for multiple comparisons, the minimum p-values for the patients are: for P1:p=0.0026; P2:p=0.0317;P3:p>0.05;P4:p=0.0109.

2.2.2 Classification on preprocessed data

To estimate classification accuracy on this data, a 10-fold crossvalidation was performed for each day and patient. A SVM with a linear kernel was used. A parameter optimization was performed inside of the cross-validation. The classification accuracy averaged over all days and patients is 49.6 %. Detailed results can be found in A.7

2.2.2 Classification on preprocessed data with slope

To classify on the slope, the slope for each 1s-window was calculated resulting in 15 (or 10) values per channel. To estimate classification accuracy on this data a 10-fold crossvalidation was performed for each day and patient. A SVM with a linear kernel was used. A parameter optimization was performed inside of the cross-validation. The classification accuracy averaged over all days and patients is 52.54 %, but none of the sessions has a classification accuracy significantly above chance level ($p<0.05$). Detailed results can be found in A.8

2.2.3 Discussion of preprocessed data

The results on the preprocessed data confirm the results obtained on the raw data and show that using slope does not lead to a significant classification accuracy. In summary, no matter if the raw data or preprocessed data was used, it was no possible to show a significant difference between yes/no responses. Classification on the raw, preprocessed or slope data also showed to be around chance-level.

3. Online Feedback and Open Session results from Chaudhary et al. (2017)

Some things about the results presented from online feedback/open session do not add up. The paper from Chaudhary et al. (2017) does not properly separate online and offline results in the methods section, so as a reader I cannot be sure about the exact methods used and if the reported results are really online results.

Based on the data that was made available to me by Ujwal Chaudhary, the results obtained online are not well organized and seem to be missing for most online/open questions sessions. In the following I will briefly summarize my impression with a focus on open question sessions; I did not check all online feedback sessions:

Patient 1:
results from open questions session are missing. With exception of P1V2D6B5
I assumed the online results to be in: ALS_BCI2014\Patient1\mat\visit2\20-M_result_f2_b5.txt and the corresponding questions in ALS_BCI2014\Patient1\mat\visit2\list of sentences\Visit2D6b5.txt

The results say that all (100%) questions were answered online with yes. According to Visit2D6b5.txt (see A.5 and A.6), there are 6 yes/no pairs. So if all are were answered with yes, this means that a maximum accuracy of 70% is possible, but a 90% accuracy is reported in the paper.

No information about visit 3 open sessions

Patient 2:

The open questions for this patient are labeled with 1 and 2, which is normally used for positive/negative questions, if I see this correctly. For Patient 1 open questions were labeled with 3. Labeling open questions as 1 or 2 might cause problems in later analysis if you are not careful in the later analysis.

For the following files, the accuracy I get does not match the results presented in the paper

Day4b4.txt & 15-A_result_f2_b4.txt : i get accuracy of 70%, paper says 85%

Day5b4 && 16-A_result_f2_b4.txt: i get accuracy of 50%, paper says 70%

Patient 3:

I couldn't find any information about feedback/open sessions

I also got an Excel Sheet from Ujwal Chaudhary containing accuracies of online/open sessions. The accuracy of 100% in all open sessions does not match the paper and seems really dubious. Possible cause of error could be the labeling. If they are labeled as open sessions instead of yes or no (which is the case in Visit2D6b5.txt), this would be classification problem of only 1 class leading to an accuracy of always 100%.

3.1 Summary

Based on the information I have, the results from the online sessions are badly documented and for the sessions where I can reproduce the results, they don't match the results published in the paper. For me, it seems that the paper published offline results based on the data of the online sessions. The methods section in the paper is not 100% clear about this.

I don't have a confirmation that the txt-files I used are actually the results of the online sessions. So this should be checked if this really are the files with the online results, or if these are stored somewhere else where I don't have access to.

As all sessions were filmed the results from the txt-files should be checked with the recorded films and also the results from the feedback sessions should be validated with the films.

4. Discussion

With the three issues outlined above, I strongly question the claim that there is NIRS-based communication in CLIS. Point 1 in itself is no evidence because it is mainly speculation about data I don't have and did not analyze, but in light of the other 2 issues it seems to me that the information about the trigger was either not fully understood and/or not properly communicated to the other people working with NIRS/CLIS.
Point 2 is the strongest argument. If there would be any effect, it should be possible to reproduce this effect, which I was not able to. Neither in a statistical analysis nor when doing a classification.
Point 3 is not directly evidence that it does not work, but only that the results presented in the paper are inaccurate. As the data, that was made available to me, is a bit unorganized with regards to the online results, it makes it difficult to check the online results, especially since information about some online sessions is completely missing.
Considering all 3 points together, I don't see any evidence that shows a working communication in CLIS patients.

However, as I also pointed out that there are some things I am not sure about or am lacking information, so I don't think that there is currently enough evidence to definitely say communication is not working. Therefore, I recommend to further investigate this issue.

4.1 Recommended next steps

To further investigate this, I recommend the following steps:

Test the online system in healthy subjects
If the system works for CLIS patients, it should work for healthy subjects. The yes/no NIRS paradigm is a completely new BCI paradigm. Any other BCI paradigm that has been published has been tested on healthy subjects. Yet, there is no publication that shows the yes/no paradigm to work in healthies.
As the online system is ready, it should be top priority to test healthy subjects and this can be done very quickly.

Reanalyze the data with a proper parameter optimization
I will make the matlab scripts I wrote available, so that my results can be reproduced. But I recommend that a third person tries to reproduce it independently. As Allesandro Tonin is already working on the classification of NIRS data, I think he should be able to do this very quickly and either reproduce or falsify the results I obtained. When doing this, I recommend to start the analysis-scripts completely from scratch and not use previous scripts to be independent of previous (possibly erroneous) work.

Check the 2013 data
As mentioned, some of the 2013 NIRS data was affected by a problem with the trigger. I don't know for sure, but it seems to me that the affected data was recorded with the Hitachi system (used for the 2014 publication) and the data recorded with the NIRX system (2014 and afterwards) is not affected by this. Therefore all data recorded with the Hitachi system should be checked for this issue or at least all data used for the 2014 publication.

Check online results
As there seem to be some mistakes in the online results, these should be double-checked and if necessary the videos of the patients should be used to obtain the real online results. Also I am not sure, if the file I used are really the online results. So, this should be checked.

Do a statistical analysis of the data
If there is an effect that can be classified, one should be able to show this effect also by a statistical analysis. Chaudhary et al. (2017) mention such an analysis, but do not show enough details about it. In the Graz conference publication Malekshahi et al. (2017) a statistical analysis was also presented. Here, a model was trained that used 4 regressors to predict O2Hb. The method used here is also questionable and should be discussed separately.

Appendix

A.1 Mail to Bin Xia regarding baseline reset

On Mon, Sep 16, 2013 at 2:16 PM, Martin Spüler <spueler@informatik.uni-tuebingen.de> wrote:
Dear Bin,

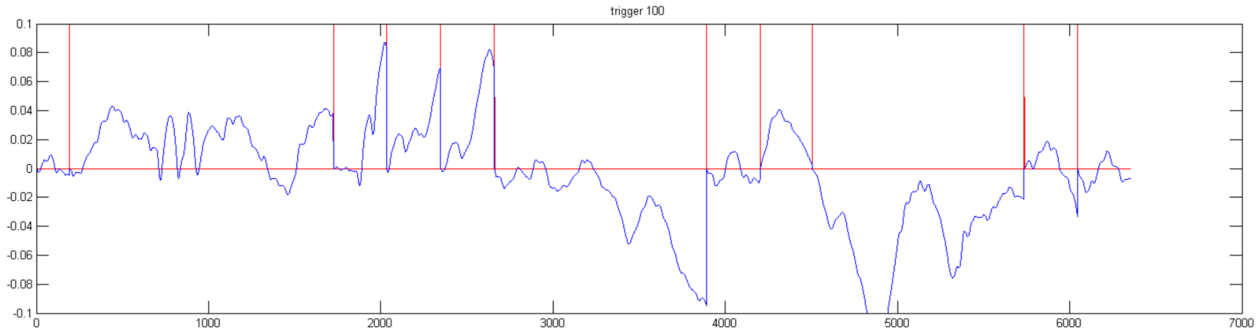
today I had a look at the raw data and as I see it, there seems to be a problem with trigger 100. Everytime trigger 100 appears, all channels are reset to 0. This does not happen with trigger 4700. This reset can cause serious problems in later spectral estimation (especially if you use some high- or low-pass filter before spectral estimation). In any case, for me it seems that this reset is a serious influence for the classification and i would not use this data for classification. In my opinion you do not classify if there is a hemodynamical response, but only if there way a reset or not.
I suggest to find out why this reset happens, fix it and redo the experiments without this reset.

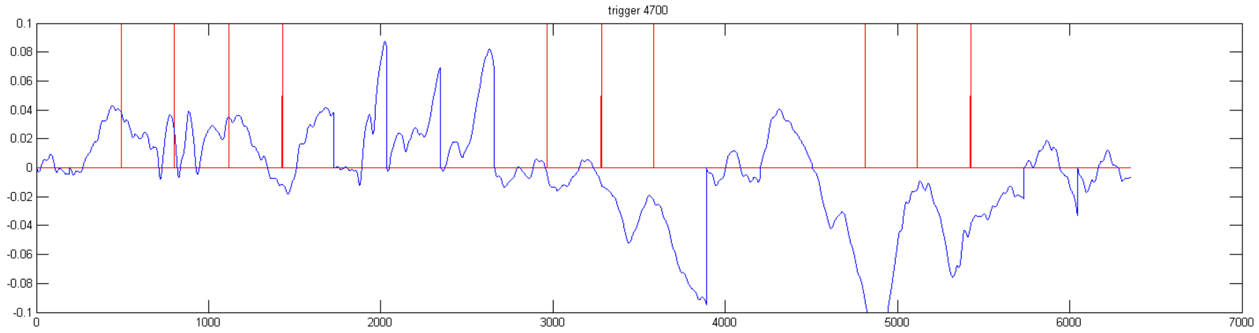
I attached you an image to show you the reset i mean. In both plots is the same dataset in blue. The red line is trigger 100 on the top plot and trigger 4700 in the bottom plot. As you can see the blue line always jumps to 0 when trigger 100 happens, but does not jump when trigger 4700 happens.

Do you have any technical information regarding this reset, why it happens (and what exactly does happen?) If this reset is only due the the amplifier adjusting some offset, this data might still be useable if you correct for the reset. If this reset happens because of some internal filtering of the NIRS-machine or something else, then you have to re-record the data.

Hope this helps and sorry that my message is unfortunate :(If you have any more questions, feel free to ask.

Best,
Martin





A.2 matlab-function used for Chaudhary et al. (2017) implementing grid-search for parameter optimization

```
function  [classifier_model]=trainSVMlinearclassifier(traindata,trainlabel,fold);
classifier_model= struct;
[traindata,classifier_model.MaxV,classifier_model.MinV] = Scale(traindata);

classifier_model.trainaccuracy=0;
%%%%%%%%%%用于选择最好的模型的参数%%%%%%%%%
for i=-2:10
    for k = -10:1:10
        tempstrc=num2str(double(2^i));
        tempstrg=num2str(double(2^k));
        fold=num2str(fold);
        sss=['-c ' tempstrc  ' -g ' tempstrg ' -v ' fold '-t 0'];  % 设置C-SVC, e -SVR和v-SVR的参数(损失函数)(默认1) n-fold交互检验模式·n为fold的个数, 必须大于等于2
%         sss=['-c ' tempstrc  ' -v ' fold];  % 设置C-SVC, e -SVR和v-SVR的参数(损失函数)(默认1) n-fold交互检验模式·n为fold的个数, 必须大于等于2

        tempaccuracy= svmtrain(trainlabel,traindata,sss);%%%%svmtrain已有函数

        if tempaccuracy>classifier_model.trainaccuracy
            classifier_model.trainaccuracy=tempaccuracy;
            strc=tempstrc;
            strg=tempstrg;
            ss=['-c ' strc ' -g ' strg  ' -b 1 -t 0'];
%             ss=['-c ' strc ' -b 1'];
            classifier_model.model = svmtrain(trainlabel,traindata,ss);%%有概率输出的模型
        end;

    end


end


fprintf('Max training data Crossaccuracy is %d \n',classifier_model.trainaccuracy);

end
```

A.3 Classification accuracy with C=1 on raw data

patient	visit	day	daystr	acc	trialcount	conflevel
1	1	1	24.03.2014	0.6	100	0.58041
1	1	2	25.03.2014	0.48333	60	0.6025
1	1	3	26.03.2014	0.35	40	0.62362
1	1	4	27.03.2014	0.53788	132	0.57031
1	1	5	28.03.2014	0.5	100	0.58041
1	2	1	15.05.2014	0.51	100	0.58041
1	2	2	16.05.2014	0.41	100	0.58041
1	2	3	17.05.2014	0.49167	120	0.57364
1	2	4	18.05.2014	0.53	100	0.58041
1	2	5	19.05.2014	0.51	100	0.58041
1	2	6	20.05.2014	0.525	80	0.58947
1	3	1	04.08.2014	0.54098	61	0.60171
1	3	2	05.08.2014	0.5875	80	0.58947
1	3	3	06.08.2014	0.49254	67	0.59732
1	3	4	07.08.2014	0.45	80	0.58947
1	4	2	05.11.2014	0.45	80	0.58947
1	4	4	07.11.2014	0.42623	61	0.60171
1	5	1	18.05.2015	0.54321	81	0.58894
1	5	2	19.05.2015	0.5	66	0.59801
1	5	3	20.05.2015	0.45	60	0.6025
1	6	2	02.09.2015	0.5125	80	0.58947
1	6	3	03.09.2015	0.48333	60	0.6025
1	6	4	04.09.2015	0.47368	38	0.62653
2	1	1	10.06.2014	0.53333	105	0.57854
2	1	2	11.06.2014	0.55	80	0.58947
2	1	3	12.06.2014	0.525	40	0.62362
2	2	1	12.08.2014	0.6	60	0.6025
2	2	2	13.08.2014	0.5	80	0.58947
2	2	3	14.08.2014	0.4375	80	0.58947
2	2	4	15.08.2014	0.59	100	0.58041
2	2	5	16.08.2014	0.59	100	0.58041
2	3	1	22.06.2015	0.38333	60	0.6025
2	3	2	23.06.2015	0.45	80	0.58947
2	3	3	24.06.2015	0.48276	58	0.60414
2	3	4	25.06.2015	0.44898	98	0.58119
2	3	5	26.06.2015	0.46	100	0.58041
2	4	1	21.09.2015	0.5125	80	0.58947
2	4	2	22.09.2015	0.60606	33	0.63481
2	4	3	23.09.2015	0.4375	80	0.58947
2	4	4	24.09.2015	0.55738	61	0.60171
3	1	1	17.06.2014	0.42308	52	0.60958
3	1	2	18.06.2014	0.45	40	0.62362
3	1	3	19.06.2014	0.52817	142	0.56786
3	1	4	20.06.2014	0.6	50	0.61159
3	1	5	21.06.2014	0.525	160	0.56403
3	2	1	25.08.2014	0.4875	80	0.58947
3	2	2	26.08.2014	0.325	40	0.62362
3	2	3	27.08.2014	0.4	60	0.6025
3	2	4	28.08.2014	0.475	40	0.62362
3	2	5	29.08.2014	0.475	80	0.58947
3	2	6	30.08.2014	0.4375	80	0.58947
3	2	7	31.08.2014	0.46667	60	0.6025
3	3	1	21.09.2014	0.6	60	0.6025
3	3	2	22.09.2014	0.55	80	0.58947
3	3	3	23.09.2014	0.525	80	0.58947
3	3	4	24.09.2014	0.4	60	0.6025
3	3	5	25.09.2014	0.56667	60	0.6025
3	3	6	26.09.2014	0.39683	63	0.60018
3	4	1	01.05.2015	0.5082	61	0.60171
4	1	1	03.09.2014	0.45	60	0.6025
4	1	2	04.09.2014	0.475	40	0.62362
4	1	3	05.09.2014	0.475	80	0.58947
4	1	4	06.09.2014	0.55	120	0.57364
4	1	5	07.09.2014	0.53333	60	0.6025
4	1	6	08.09.2014	0.475	40	0.62362
4	2	1	15.12.2014	0.475	40	0.62362
4	2	2	16.12.2014	0.48	100	0.58041
4	2	3	17.12.2014	0.575	80	0.58947
4	2	4	18.12.2014	0.50877	57	0.60499
4	2	5	19.12.2014	0.61667	60	0.6025
4	3	1	20.01.2015	0.48333	120	0.57364

A.4 Classification accuracy with parameter optimization on raw data

patient	visit	day	daystr	acc	trialcount	conflevel
1	1	1	24.03.2014	0.61	100	0.58041
1	1	2	25.03.2014	0.41667	60	0.6025
1	1	3	26.03.2014	0.375	40	0.62362
1	1	4	27.03.2014	0.40909	132	0.57031
1	1	5	28.03.2014	0.46	100	0.58041
1	2	1	15.05.2014	0.52	100	0.58041
1	2	2	16.05.2014	0.52	100	0.58041
1	2	3	17.05.2014	0.45833	120	0.57364
1	2	4	18.05.2014	0.44	100	0.58041
1	2	5	19.05.2014	0.43	100	0.58041
1	2	6	20.05.2014	0.65	80	0.58947
1	3	1	04.08.2014	0.4918	61	0.60171
1	3	2	05.08.2014	0.55	80	0.58947
1	3	3	06.08.2014	0.46269	67	0.59732
1	3	4	07.08.2014	0.3875	80	0.58947
1	4	2	05.11.2014	0.5	80	0.58947
1	4	4	07.11.2014	0.60656	61	0.60171
1	5	1	18.05.2015	0.55556	81	0.58894
1	5	2	19.05.2015	0.5	66	0.59801
1	5	3	20.05.2015	0.51667	60	0.6025
1	6	2	02.09.2015	0.6	80	0.58947
1	6	3	03.09.2015	0.31667	60	0.6025
1	6	4	04.09.2015	0.57895	38	0.62653
2	1	1	10.06.2014	0.42857	105	0.57854
2	1	2	11.06.2014	0.55	80	0.58947
2	1	3	12.06.2014	0.475	40	0.62362
2	2	1	12.08.2014	0.58333	60	0.6025
2	2	2	13.08.2014	0.525	80	0.58947
2	2	3	14.08.2014	0.3375	80	0.58947
2	2	4	15.08.2014	0.51	100	0.58041
2	2	5	16.08.2014	0.55	100	0.58041
2	3	1	22.06.2015	0.51667	60	0.6025
2	3	2	23.06.2015	0.4875	80	0.58947
2	3	3	24.06.2015	0.5	58	0.60414
2	3	4	25.06.2015	0.54082	98	0.58119
2	3	5	26.06.2015	0.53	100	0.58041
2	4	1	21.09.2015	0.5875	80	0.58947
2	4	2	22.09.2015	0.60606	33	0.63481
2	4	3	23.09.2015	0.4375	80	0.58947
2	4	4	24.09.2015	0.57377	61	0.60171
3	1	1	17.06.2014	0.53846	52	0.60958
3	1	2	18.06.2014	0.4	40	0.62362
3	1	3	19.06.2014	0.47887	142	0.56786
3	1	4	20.06.2014	0.5	50	0.61159
3	1	5	21.06.2014	0.5125	160	0.56403
3	2	1	25.08.2014	0.4875	80	0.58947
3	2	2	26.08.2014	0.5	40	0.62362
3	2	3	27.08.2014	0.48333	60	0.6025
3	2	4	28.08.2014	0.375	40	0.62362
3	2	5	29.08.2014	0.3875	80	0.58947
3	2	6	30.08.2014	0.5125	80	0.58947
3	2	7	31.08.2014	0.36667	60	0.6025
3	3	1	21.09.2014	0.51667	60	0.6025
3	3	2	22.09.2014	0.4375	80	0.58947
3	3	3	23.09.2014	0.45	80	0.58947
3	3	4	24.09.2014	0.55	60	0.6025
3	3	5	25.09.2014	0.51667	60	0.6025
3	3	6	26.09.2014	0.39683	63	0.60018
3	4	1	01.05.2015	0.47541	61	0.60171
4	1	1	03.09.2014	0.53333	60	0.6025
4	1	2	04.09.2014	0.475	40	0.62362
4	1	3	05.09.2014	0.6125	80	0.58947
4	1	4	06.09.2014	0.53333	120	0.57364
4	1	5	07.09.2014	0.61667	60	0.6025
4	1	6	08.09.2014	0.6	40	0.62362
4	2	1	15.12.2014	0.5	40	0.62362
4	2	2	16.12.2014	0.54	100	0.58041
4	2	3	17.12.2014	0.5875	80	0.58947
4	2	4	18.12.2014	0.54386	57	0.60499
4	2	5	19.12.2014	0.5	60	0.6025
4	3	1	20.01.2015	0.51667	120	0.57364

A.5 content of *20-M_result_f2_b5.txt*

```
1
1
1
1
1
1
1
1
1
1
1
1
1
1
1
1
1
1
1
```

A.6 content of *Visit2D6b5.txt:* sentences with positive/negative pair are marked with XX

```
003_1220.wav
2
003_2203.wav XX
2
003_2218.wav
2
003_1208.wav
2
003_1254.wav
2
003_1252.wav
2
003_1255.wav
2
003_2202.wav XX
2
003_1215.wav
2
003_1235.wav
2
003_1227.wav XX
2
003_1202.wav XX
2
003_2207.wav XX
2
003_1232.wav
2
003_2219.wav XX
2
003_1207.wav XX
2
003_2227.wav XX
2
003_1219.wav XX
2
003_1203.wav XX
2
003_2212.wav
2
```

A.7 Clasification results on preprocessed data: First row is accuracy, second confidence level, third if it is significant with p<0.05; each column is one day. Note that the preprocessed data did not contain all days for Patient 1.

Fig2A+Table1_PatientF_P1_Ham

0.484	0.5325	0.5275	0.4963	0.489	0.445
0.5804	0.5895	0.5736	0.5895	0.5804	0.5895
0	0	0	0	0	0

Fig2B+Table1_PatientG_P3_Bai

0.5813	0.46	0.4845	0.3523	0.5056	0.4133	0.5544	0.485	0.415	0.53	0.4383	0.5475	0.4725	0.4725	0.4825	0.4525	0.425	0.5775
0.6367	0.6674	0.5679	0.6184	0.564	0.6025	0.59	0.6025	0.6025	0.6025	0.6025	0.5895	0.6236	0.6236	0.6236	0.5895	0.6025	0.6236
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Fig2C+Table1_PatientB_P2_Gri

0.54	0.5563	0.425	0.4788	0.4162	0.5183	0.5175	0.4383	0.5612	0.5363	0.4387	0.5188
0.5895	0.5895	0.6236	0.5895	0.5895	0.6025	0.5895	0.6025	0.5895	0.5895	0.5895	0.5895
0	0	0	0	0	0	0	0	0	0	0	0

Fig2D+Table1_PatientW_P4_Lev

0.58	0.52	0.52	0.4	0.515	0.495	0.5	0.585	0.445	0.495	0.525	0.545	0.46	0.495	0.39	0.435	0.435	0.7	0.705	0.515
0.6674	0.6674	0.6674	0.6674	0.6674	0.6674	0.6674	0.6674	0.6674	0.6674	0.6674	0.6674	0.6674	0.6674	0.6674	0.6674	0.6674	0.6674	0.6674	0.6674
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0

A.8 Clasification results on preprocessed data with slope: First row is accuracy, second confidence level, third if it is significant with $p<0.05$; each column is one day. Note that the preprocessed data did not contain all days for Patient 1.

Fig2A+Table1_PatientF_P1_Ham

0.528	0.5262	0.53	0.53	0.527	0.515
0.5804	0.5895	0.5736	0.5895	0.5804	0.5895
0	0	0	0	0	0

Fig2B+Table1_PatientG_P3_Bai

0.5225	0.5275	0.515	0.5275	0.5262	0.5267	0.535	0.5267	0.5375	0.5275	0.5212	0.5313	0.5225	0.5275	0.515	0.5275	0.5262	0.5267
0.5895	0.5895	0.6236	0.5895	0.5895	0.6025	0.5895	0.6025	0.5895	0.5895	0.5895	0.5895	0.5895	0.5895	0.6236	0.5895	0.5895	0.6025
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Fig2C+Table1_PatientB_P2_Gri

0.5281	0.525	0.5201	0.5023	0.5212	0.5117	0.5282	0.5183	0.52	0.53	0.5217	0.5212
0.6367	0.6674	0.5679	0.6184	0.564	0.6025	0.59	0.6025	0.6025	0.6025	0.6025	0.5895
0	0	0	0	0	0	0	0	0	0	0	0

Fig2D+Table1_PatientW_P4_Lev

0.54	0.515	0.525	0.535	0.52	0.54	0.535	0.55	0.53	0.54	0.515	0.52	0.505	0.52	0.525	0.515	0.52	0.53	0.545	0.52
0.6674	0.6674	0.6674	0.6674	0.6674	0.6674	0.6674	0.6674	0.6674	0.6674	0.6674	0.6674	0.6674	0.6674	0.6674	0.6674	0.6674	0.6674	0.6674	0.6674
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0